Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection Yong Pung How School Of Law

Yong Pung How School of Law

4-2024

Discovering significant topics from legal decisions with selective inference

Jerrold Tsin Howe SOH
Singapore Management University, jerroldsoh@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sol_research

Part of the Courts Commons, Legal Studies Commons, and the Numerical Analysis and Scientific Computing Commons

Citation

SOH, Jerrold Tsin Howe. Discovering significant topics from legal decisions with selective inference. (2024). *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 382, (2270), 1-21.
Available at: https://ink.library.smu.edu.sg/sol_research/4433

This Journal Article is brought to you for free and open access by the Yong Pung How School of Law at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection Yong Pung How School Of Law by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

PHILOSOPHICAL
TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Check for updates

Cite this article: Soh Tsin Howe J. 2024 Discovering significant topics from legal decisions with selective inference. *Phil. Trans. R. Soc. A* **382**: 20230147. https://doi.org/10.1098/rsta.2023.0147

Received: 9 July 2023
Accepted: 19 December 2023

One contribution of 15 to a theme issue 'A complexity science approach to law and governance'.

Subject Areas:
artificial intelligence

Keywords:
legal language processing, topic models, text-as-data, domain name disputes, European Court of Human Rights

Author for correspondence:
Jerrold Soh Tsin Howe
e-mail: jerroldsoh@smu.edu.sg

THE ROYAL SOCIETY
PUBLISHING

Discovering significant topics from legal decisions with selective inference

Jerrold Soh Tsin Howe

Yong Pung How School of Law, Singapore Management University School, Singapore

JSTH, 0000-0003-0270-5015

We propose and evaluate an automated pipeline for discovering significant topics from legal decision texts by passing features synthesized with topic models through penalized regressions and post-selection significance tests. The method identifies case topics significantly correlated with outcomes, topic-word distributions which can be manually interpreted to gain insights about significant topics, and case-topic weights which can be used to identify representative cases for each topic. We demonstrate the method on a new dataset of domain name disputes and a canonical dataset of European Court of Human Rights violation cases. Topic models based on latent semantic analysis as well as language model embeddings are evaluated. We show that topics derived by the pipeline are consistent with legal doctrines in both areas and can be useful in other related legal analysis tasks.

This article is part of the theme issue 'A complexity science approach to law and governance'.

1. Introduction

Most legal information is stored exclusively in natural language texts. The complexity of language means extracting such information is typically a labour-intensive exercise primarily performed by specially trained persons (lawyers). This poses significant barriers to computational representation and analysis of law [1,2]. Researchers have increasingly sought to develop automated processes for converting unstructured legal texts to structured variables [3,4]. Depending on the texts involved and variables required, these have included term frequency counts [5], regular expressions [6],

© 2024 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

topic models [7–11], word embeddings [12,13] and language models [14–16]. Given their centrality in legal analysis, court decisions in particular have attracted significant scholarly attention. Many studies have attempted to identify, categorize or forecast case outcomes using decision texts, often relying on opaque algorithms such as support vector machines and neural networks [7,17–20]. Other researchers have prioritized more explainable methods over end-to-end prediction. Typically, algorithms are first developed to automatically extract case attributes and other legally relevant variables before using these variables to model outcomes [8,21–24]. The goal is not necessarily predictive accuracy alone, but also to identify and explain what motivates legal decisions.

In this work, we propose and evaluate a new automated pipeline for discovering significant topics from decision texts, a task we define more formally in §2a. The pipeline takes decision texts and case outcomes as inputs and returns estimates for statistically significant decision topics as well as the cases, words and phrases most strongly associated those topics. This allows researchers to quickly identify potential variables, patterns and cases of interest in unfamiliar areas of law. The pipeline comprises four steps: pre-processing and masking (§2b(i)), topic modelling (§2b(ii)), selective regression and inference (§2b(iii)) and topic evaluation (§2b(iv)). We demonstrate and evaluate the pipeline on a new dataset of cases resolved under the Uniform Domain Name Dispute Resolution Policy (UDRP). To explore how the pipeline generalizes, we further test it on a canonical dataset of European Convention of Human Rights (ECHR) cases. For both datasets, we experiment with latent semantic analysis (LSA) [25] as well as two BERTopic (BTO) models [26] primed with general and legally fine-tuned embeddings, respectively.

We show that topics discovered by the pipeline contain interpretable and legally sound information on case patterns correlated with legal outcomes (§3). Along the way, we identify several interesting patterns and case archetypes in UDRP and ECHR case law. Thus, our key contributions are as follows. First, we extend prior work analysing legal outcomes from a topic modelling perspective [7,8,11]. To be sure, the notion that topics synthesized from case decisions could carry meaningful information about legal outcomes is not new. Neither do we propose entirely new algorithms for, say, legal topic modelling. Our incremental contribution lies in integrating several existing techniques (e.g. masking [27], topic modelling [7,11] and selective inference [28,29]) into a pipeline that can be adapted to study other legal areas. Second, we demonstrate the utility of selective inference techniques in the legal domain. This has not, to our best knowledge, been studied in prior work. Finally, we add to legal knowledge on UDRP and ECHR cases.

2. Methods

(a) Discovering significant topics

This work relates to existing literature on the automated extraction of legal factors from legal cases [8,23,24]. Legal factors are generally seen as ‘stereotypical patterns of fact’ [8] or more abstract ‘intermediate concepts’ [30] which influence case outcomes. However, as used in that literature, the concept of legal factors has a specific meaning which does not overlap perfectly with our present focus. We thus use the term ‘predictors’ here to refer broadly to variables which predict case outcomes. Drawing inspiration from [31], suppose a legal outcome Y is given by $Y = f(X, W)$, where X is a matrix of legal predictors, W a matrix of non-legal predictors (e.g. political ideologies [32]), and f some adjudication function that maps cases to outcomes. To identify individual predictors, we might collect data on hypothesized variables $\hat{X}$, $\hat{W}$ (i.e. approximations of X and W), and estimate the model $\hat{Y} = \hat{f}(\hat{X}, \hat{W})$. Weights computed for each $\hat{x}, \hat{w}$ would capture the strength and polarity of their correlation with outcomes. Variables assigned significant, non-zero weights can be understood as potential legal (or non-legal) predictors. They may further be seen as *causal* predictors, if the model is causally identified, or correlative predictors otherwise.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 1. Tasks involving legal predictors. This work focuses on the discovery task.

given inputs	goal	task label(s)
$\hat{X}, \hat{W}$	estimate weights	analysis
D, candidate $\hat{X}, \hat{W}$s known	extract observations of $\hat{X}, \hat{W}$	identification/extraction
D, candidate $\hat{X}, \hat{W}$s unknown	discover candidate $\hat{X}, \hat{W}$s	discovery

The challenge with legal applications is that the x's and w's are not available as structured data but found only in some natural language corpus D. Typically, these are decision texts written to state and justify outcomes for each case i, though other documents including submissions, affidavits and procedural records may also be relevant. We must apply a 'codebook function' $g : D \mapsto Q$, $Q \in \mathbb{R}^{n\times m}$ that maps n texts to m variables [3]. Where the variables desired are known *ex ante* based on legal domain knowledge, the researcher's aim is to extract *observations* of $\hat{x}_i$. But in unfamiliar legal areas where candidate predictors are not already known, the goal shifts from filling observations or estimating coefficients to *discovering* such predictors to begin with. There are therefore three different tasks related to legal predictors (table 1). Notably, these are not mutually exclusive and must often be performed in tandem to answer the research question. Suppose as in [33] that we want to know if case origin influences the probability of a certiorari grant by the US Supreme Court. The variable of interest is known but potential confounders remain to be identified. We would need to extract observations for case origin, discover (and thereafter extract observations for) potential confounders and finally analyse coefficients for case origin while controlling for these confounders. This work is chiefly concerned with the discovery task, though extraction and analysis are by-products of the proposed method.

(b) Proposed pipeline

(i) Pre-processing and masking

We begin with a text corpus D and structured categorical outcomes Y for n cases in some legal area of interest. In theory, any corpus with sufficient case information, such as case briefs and affidavits, could be used. In practice, most legal analysis is based solely on decision texts. Other legal documents are usually not accessible at scale. Thus, we tailor the approach assuming D is a decision corpus. The use of decision texts has important implications for the kind of analyses possible and the pre-processing steps necessary. Specifically, fitting legal outcome models on decisions is problematic because decisions are written by judges, after observing case facts, to justify case outcomes [33]. Extracted features could therefore contain both post-treatment and post-outcome information, making them 'bad controls' [34]. Formally, suppose decisions are generated by the process $D = t(Y, X, W, J)$, where J accounts for the judges' individual writing styles and t is some text-generation function. Substituting this into the model $\hat{Y} = g(D)$ gives $\hat{Y} = g(t(Y, X, W, J))$. Since we are indirectly modelling Y on itself, we should expect the model to produce large, significant estimates for features still containing hints of Y after the transformations g and t instead of unbiased estimates for x's and w's.

As we do not control t, the natural solution, other than switching to some pre-outcome corpus, is to build into g processes for masking information on Y. We follow standard steps from the legal prediction literature in masking outcome-revealing sections of and phrases in the text from the model by deleting them entirely at the start of the pipeline [7,27]. This may over-inclusively remove otherwise informative words, but is however taken as a necessary and non-fatal trade-off [27]. It also may not remove all outcome information from D. Since decisions are written to *justify* case outcomes, even seemingly innocuous sections such as 'case facts' could be arranged in a way that favours the writer's preferred outcome. Indeed, lawyers are typically taught to present facts *persuasively* [35]. This pertains especially to case briefs, but we cannot preclude its occurrence in

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

decisions. As such, we emphasize that predictors discovered by our method should be interpreted as *correlative*.

Where required, we then pre-process the masked corpus in standard fashion by lowercasing, stopping, and lemmatization. This applies mainly to LSA as BTO is trained on raw texts.[1]

(ii) Topic modelling

Topic models are suitable codebooks because of their readability: each $q \in Q$ can be manually interpreted based on representative n-grams, and documents with higher q weights can be read as being more heavily or likely 'about' q. Of the numerous topic models in the literature, here we experiment with one hot encoding (i.e. indicators for each n-gram in the corpus overall vocabulary) (OHE), LSA and BTO to cover a range of traditional and emerging approaches. As topic models are well documented elsewhere, below we provide a condensed description of those we test.

LSA first computes a term-frequency/inverse-document-frequency (TFIDF) encoding [36–38]. The TFIDF matrix is compressed into m desired topics (explained below) by applying singular value decomposition (SVD) and keeping only features corresponding to the largest m singular values. The SVD of a matrix is $W = U_m S_m \Lambda_m^T$, where $\text{rank}(W) = m$, $m \leq \text{rank}(W)$ [39]. When W is a TFIDF matrix, U_m corresponds to n-gram vectors, S_m to singular values of W, and Λ_m to document vectors [38]. The corpus is thus represented through Λ_m as a distribution of m topics across n documents [25]. These 'topics' are represented in U_m as distributions across n-grams. For intuition, observe that an optimal compression of term frequency matrices should squeeze co-informative terms together, forming said topics. We use LSA here because of its prominence in the influential work of Aletras *et al*. [7] on legal outcome prediction for ECHR cases as well as subsequent related work.

BTO [26] is modular framework which starts with paragraph embeddings typically derived from a language model. Depending on the LM's context window, longer documents may be partitioned into smaller chunks if necessary [40]. Chunk embeddings undergo dimensionality reduction via a standard algorithm such as UMAP [41] (the default) or principal components analysis before clustering via another algorithm such as HDBSCAN [42] or k-Means. Topics are extracted from these clusters using a bag-of-words vectorizer followed by a 'class-based' TFIDF implementation given by $cTFIDF(c) = ||tf_{w,c}|| \times \log(1 + (A/f_w))$, where $tf_{w,c}$ is the frequency of n-gram x in cluster c, f_w is the frequency of w's frequency across all clusters, and A is the average number of tokens per cluster. This produces an arbitrary number of topics which can be iteratively merged based on topic frequency and *cTFIDF* similarity until a desired number remains. The resulting chunk-topic matrix can then be re-constituted into document-level topics in several ways. For instance, by assigning a document to the one topic which contains the largest number of its chunks (i.e. max-pooling). Following [40], we take chunk-topic counts normalized at document level. We test two BTO models primed with chunk embeddings from (i) all-MiniLM-L6-v2 [43], a sentence transformer based on [44] and recommended by Grootendorst [26] (BTO$_M$) and (ii) legalBERT [14], a BERT [45] extension fine-tuned on UK, EU and US legal documents (BTO$_L$). Inspiration for using BTO in the legal context comes from [11] which used a multilingual MiniLM-embedded BTO model to study Canadian housing law court decisions written in French.

Here we generate topics comprising $1, 2, 3$-grams for all topic models. For LSA, we generally take only the 2500 most frequent n-grams at the TFIDF step before reducing the matrix to a desired topic number based on corpus size. As context, the best predictive models of Aletras *et al.* [7] for ECHR cases generally used LSA topics creating with the 2000 top $1, 2, 3, 4$-grams. However, in our (unreported) exploratory tests, we noted that 4-grams do not add new interpretable information as they usually repeated terms already seen in $1, 2, 3$-grams. We also set minimum document frequency cutoffs of 5 or 10 (depending on dataset and topic model) in LSA's TFIDF and BTO's cTFIDF steps to limit computational and memory overheads. Other parameters follow recommendations and defaults from the `sklearn` [46] and `bertopic` [26] libraries.

[1] https://maartengr.github.io/BERTopic/faq.html#how-do-i-remove-stop-words.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 2. UDRP summary statistics by outcome. Mean values presented. Standard deviations in parentheses. Raw word count includes all tokens in the text after removing only the 'Decision' section. Processed word count includes only tokens remaining after lower-casing, stopword removal and lemmatization were further applied.

	complainant won	complainant lost	overall
no. comp'ts	1.108	1.095	1.106
	(0.393)	(0.392)	(0.393)
no. resp'ts	1.094	1.077	1.092
	(1.301)	(0.349)	(1.232)
no. domain names	2.009	1.407	1.942
	(14.749)	(2.188)	(13.921)
raw word count	2812.439	3560.197	2895.986
	(1309.640)	(1680.681)	(1376.408)
processed word count	1293.598	1630.689	1331.261
	(600.045)	(772.365)	(630.659)
GTLD cases (%)	94.32	94.94	94.39
three-member panel cases (%)	3.21	24.50	5.59
comp't won (%)	100.00	0.00	88.83
N	20 122	2531	22 653

(iii) LASSO regression and selective inference

We use a LASSO [28] regression model to associate topics with outcomes. The LASSO uses the coefficient vector's L1-norm as a penalty term when optimizing the model, such that the objective function becomes $L(\beta)^* = L(\beta) - \lambda\|\beta_j\|$, where λ is a user-specified 'shrinkage parameter' that controls penalization magnitude, and $j > 0$ (the intercept is not penalized). The LASSO is suitable for legal outcome models in three ways. First, as the goal is to discover interpretable legal topics rather than inexplicably predict legal outcomes, regression models are preferable to more opaque approaches like neural networks. Second, the LASSO overcomes two common, related problems with legal outcome models. First, as text feature matrices are typically large and sparse, and legal corpora often yield few observations, legal outcome models are prone to the $k \gg n$ problem [47,48]: as k approaches and eventually exceeds n standard regression models relying on maximum-likelihood estimation are liable to producing biased estimates or failing to converge entirely. Second, legal areas often present highly imbalanced response classes, forcing us to estimate 'rare events' [49–51]. For instance, in our UDRP dataset, approximately 90% of the cases are decided in the complainant's favour (table 2). Coupled with $k \gg n$, legal outcome models could be perfectly separated—outcomes can be perfectly predicted with a subset of features—preventing model convergence. Penalized regressions are one standard countermeasure to both problems [48,49,52–54]. In bioinformatics and chemometrics, LASSO regressions have been successfully deployed in studies involving large feature matrices and rare events [55,56].

Third, the LASSO lets us exploit emerging methods for selective inference. Conventionally, significance tests are not done with penalized regressions since regularization means estimates are biased toward zero and not consistent [48]. Nonetheless, LASSO regressions were demonstrably capable of selecting the most significant regressors, particularly in a $k \gg n$ setting [48]. More recently, [57] devised a method for conducting valid post-selection significance tests which [29] extend to the LASSO. *P*-values are computed after de-biasing the model post-selection [29,57]. Coefficient estimates must still be interpreted in light of the penalty, but *p*-values and standard errors remain valid and have been shown to be more reliable than non-adjusted values

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

from subset-selected models [48,57]. Notably, if as cautioned above we confine ourselves to discovering *correlative* rather than causal predictors, significance test validity is less of a concern. We use the Taylor & Tibshirani [29] R package `selectiveInference` [58] and following their documentation estimate the LASSO with `glmnet` [59].

(iv) Evaluation

We test several model specifications for the primary UDRP dataset, varying whether topic features are included and the topic model used (see §2c). Each specification is also evaluated on standard measures of fit including the area under the receiver operating characteristic curve (AUROC) and the median deviance ratio (MDR). The latter summarizes all deviance ratios reported by `glmnet` along the λ fitting path and can be interpreted as the pseudo-R^2 [59]. We manually evaluate selected specifications by delving into topics with the largest positive or negative coefficients and the smallest p-values for those specifications. The author, who is legally trained, then studied the topics' n-gram distributions and the cases most strongly associated with them to see how far they corresponded with topics known to be significant in legal doctrine. Notice that even if they do not, topics discovered this way could point to some yet unknown X or W driving legal outcomes. This step should therefore be informed by legal theory. To be sure, we do not suggest it can be fully automated, nor that the method is sufficient to identify all legally significant topics.

Other than evaluation, the method requires structured data in only two respects. First, labelled case outcomes are needed. While not considered in this work, existing methods for automated legal outcome extraction (e.g. [7,18,20,60]) could be incorporated at an earlier pipeline step. Second, tailored pre-processing work is necessary to sectionize documents and to mask outcome-leaking information. Other than in these two areas, topics correlated with legal outcomes are automatically synthesized from the corpus, selected by the LASSO, and surfaced by post-selection significance tests. Prior domain knowledge of potential legal predictors within the given legal area is neither assumed nor required, though it would certainly be a bonus. Likewise, while structured case metadata are not strictly needed, any available variables can easily be included as additional covariates at the regression stage.

(c) Datasets

(i) Domain name disputes

The UDRP is a mandatory policy instituted in 1999 by the Internet Corporation for Assigned Names and Numbers (ICANN) for resolving disputes over generic top-level domains (GTLDs). Several countries have adopted similar policies for their country-coded top-level domains (CCTLDs) [61]. Disputes are administered by ICANN-appointed Dispute Resolution Providers (DRPs). The largest DRP by disputes resolved is the World Intellectual Property Organization (WIPO). Under the *Rules for Uniform Domain Name Dispute Resolution Policy*, a case begins when a trademark holder files a complaint with a DRP. The DRP will ask the respondent for a written response, and thereafter assemble an adjudication panel of 1 or 3 panellists, depending on the parties' preferences. Under UDRP Article 4a, the complainant must show that (i) the contested domain is 'identical or confusingly similar' to the complainant's trade or service mark; (ii) the respondent does not have any 'rights or legitimate interests' in the contested domain; *and* the contested domain was 'registered and used in bad faith'. While parties may be represented by lawyers, all procedures are written and there are no physical hearings. If a complaint succeeds, the panel may order the domain to be transferred to the complainant or be cancelled altogether. Decisions are communicated to and enforced by the relevant domain name registrar [62].

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

We obtained from WIPO's online database[2] decision texts for WIPO-administered UDRP disputes decided on and between 1999 and 2016. Regular expressions were developed, by iterative testing on randomly sampled decisions, to partition the texts into eight archetypal sections. Case outcomes are typically stated in a final section titled 'The Decision', and occasionally in a preceding section generally titled 'Discussion and Findings'. The latter details the panel's legal reasoning and analysis. Both sections were masked. Outcome labels 'transfer', 'cancel' and 'deny' and linguistic variants thereof were also removed. This left only sections on case facts, parties involved, procedural history and arguments presented for downstream processing. Decisions where fewer than all eight sections could be detected, either because they were not in English or because of exceptional or missing headers, were excluded. This reduced the initially downloaded set of 27 634 raw cases into 22 653 usable observations.[3]

Labelled outcomes and other structured variables were extracted from case summary tables on WIPO's website. Each table contains case number, decision date, the domains, parties, and panellists involved, and outcome. While only three outcomes (i.e. transfer, cancellation or complaint denied) are possible per domain, cases with multiple domains could present mixtures (e.g. complaint denied, transfer in part with dissenting opinion). Nonetheless, the vast majority (98.87%) of cases involved singular outcomes. By studying the data, we found that outcome statements start with the outcome assigned to a majority of the contested domains (i.e. in the example above most domains would *not* have been transferred). We thus binarized outcomes by recording 1 when the outcome statement begins with 'Transfer' or 'Cancellation', and 0 when it begins with 'Complaint denied'. Basic string methods were used to extract other variables from the tables, including the number of panellists, complainants, respondents and domain names involved, whether the case involved GTLDs or CCTLDs, and year and month indicators. We also created indicators for repeat complainants (respondents) appearing in greater than 100 (30) cases.

Identity indicators were also created for all panellists. We use this to demonstrate how the method could be instrumental for studying how judge identity influences legal outcomes, a staple in 'judicial behaviour' research [63]. Legal scholars have debated the UDRP's merits [64], with critics alleging structural pro-complainant biases in the UDRP procedural rules [65–68]. Proponents [69–71] countered that critics fail to account for specific case attributes. Empirical analyses have offered different explanations for high complainant success rates. Kesan & Gallo [72] argued that case resolution efficiency was as important as apparent bias in determining provider choice, while Klerman [73] used an alternative linear regression methodology on Kesan & Gallo's [72] dataset of 2000–2001 cases to show the opposite: that complainants chose providers based on success likelihoods rather than resolution speed.

Table 2 summarizes the dataset. It contains information on more cases and variables than an earlier UDRP corpus compiled by Branting *et al.* [22]. On these data, we run the penalized logit regression:

$$\text{complainantwon}_i = \text{panelistidentity}_i + \text{panelsize}_i + \text{textfeatures}_i + \text{controls}_i + \epsilon_i,$$

where complainantwon_i is an indicator for complaint success, $\text{panelistidentity}_i$ an indicator matrix for panellist involvement, panelsize_i indicates if the case involved three panellists or one, and textfeatures_i is either an OHE, LSA or BTO document-topic matrix. controls_i are indicators for year and month, repeat player involvement, and whether the case involved GTLDs or CCTLDs. As indirect controls for dispute complexity, we also included the raw and processed word counts of the relevant decision, as well as the number of complainants, respondents, and domain names involved.

To investigate the topic models' impact, we estimate regressions with/without topic features across three settings: (A) only 1-panellist GTLD cases, (B) all GTLD cases and (C) all cases. We

[2] www.wipo.int/amc/en/domains/search/.

[3] Most of the 4981 cases dropped here were non-English cases. This is consistent with official WIPO statistics, available at https://www.wipo.int/amc/en/domains/statistics, which show that around 85–88% of cases each year are in English.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 3. Summary statistics for the ECHR dataset. Mean values presented with standard deviations in parentheses. Note that Medvedeva *et al.* [19] had balanced the dataset by random under-sampling.

	Article 3		Article 6		Article 8	
article contested:	violation	no violation	violation	no violation	violation	no violation
raw word count	4313.676	5135.169	1768.104	3894.557	4063.601	4909.45
	(3579.273)	(4046.314)	(2054.486)	(2894.177)	(4003.777)	(2897.937)
processed word count	2020.648	2424.053	821.394	1800.873	1891.228	2277.022
	(1680.586)	(1916.545)	(943.129)	(1344.95)	(1893.032)	(1353.649)
N	284	284	454	449	228	229

partition the data by panel size and domain type because these give rise to qualitatively different case types. To evaluate models in the same regression setting on similar bases, we extract exactly 250 topics with each topic model. We chose 250 after some iterative testing with LSA because it represented a 90% compression of the original TFIDF matrix (recall that the top 2500 n-grams were used) but, as computed by the SVD, explained about 61% of the variance in the same. Around the 250 mark, reducing (increasing) the number of topics led to more (less) than proportionate losses (gains) in variance explained. We used LSA rather than BTO models to experiment with topic number because re-estimating BTO models requires significantly more compute. There is some inevitable arbitrariness here as identifying the appropriate number of topics is a known challenge in topic modelling [74]. Future work could study how emerging techniques for doing so (e.g. [75,76]) could be incorporated into our method.

All topic models are trained using only decisions within the relevant partition. This except for BTO chunk *embeddings* (only the first step) which are pre-computed only once on the entire corpus and used across all settings, as the embedding process is computationally expensive. We also pre-computed the shrinkage parameter λ to be used using specifications *without* text features following the guideline suggested in [7,77] to set $\lambda = 2\mathbf{E}[\|X^T\epsilon\|_\infty]$, where $\epsilon \sim N(0, \hat{\sigma}^2)$ and $\hat{\sigma}^2$ is the residual sum of squares from a simple linear regression of y on all regressors. The same λ's were then used for mirror specifications *with* text features. As a further baseline, we also tested specifications with white noise placebos [78].

(ii) European Convention on Human Rights violations

The ECHR establishes fundamental human rights for signatory jurisdictions, including the prohibition of torture (Article 3), right to a fair trial (Article 6), and right to respect for private and family life (Article 8). The European Court of Human Rights (ECtHR) adjudicates complaints. The court publishes decision texts and 'case detail' tables on its 'HUDOC' database.[4] ECHR cases have been studied in several prior works [7,19] and included in the benchmark LexGLUE [79]. While LexGLUE provides a large number of processed ECHR texts and outcomes, that dataset is not linked to case identifiers, making topic interpretation challenging. Here we use the dataset of Medvedeva *et al*. [19] and replicate their pre-processing steps with their published code. We limit the analysis to training set cases with clear violation/non-violation outcomes (i.e. not filed in the dataset as 'both'). Below we focus on Articles 3, 6 and 8 which have the largest number of cases in this dataset. Following [19], we use only text from the Procedure, Circumstances, and Relevant Law sections. Table 3 summarizes the dataset. As our aim was to demonstrate generalizability, unlike with the UDRP we did not further extract new case variables. The main specification tested is $\text{violation}_i = \text{textfeatures}_i + \epsilon_i$ with textfeatures_i being 100 topics synthesized using the above topic models.

[4] http://hudoc.echr.coe.int/.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

3. Results

(a) UDRP results

Table 4 summarizes our primary results on the UDRP dataset. Columns 1–3 report baseline estimates computed without any text features for three main regression settings. Around 50 panellists are significant at $\alpha = 0.05$ across these baselines even with several controls included (column 3), suggesting an association between their involvement and complaint outcomes. The association is notably weaker in the corresponding topic regressions with OHE, LSA, BTO$_M$ and BTO$_L$ features added (columns 4, 5–7, 8–10 and 11–13). The topic regressions consistently yield fewer significant panellists, smaller panellist effects and better model fits. Statistical significance can be observed shifting towards the topics instead. This can already be observed with simple OHE, but is clearest with the LSA regressions, where few panellists remain significant (9, 13, 8 in columns 5–7 versus 53, 50, 49 in columns 1–3). Across all regression settings, LSA consistently produces the largest number of significant topics and the highest fit scores. Table 4, column 7 in particular yields 32 significant topics but only 8 significant panellists at $\alpha = 0.05$ and the highest MDR (0.431) and AUROC (0.914). BTO$_L$ and BTO$_M$ yield more significant panellists and fewer significant topics, but are nonetheless superior to the non-text and white noise (table 4, column 14) baselines, suggesting that these topic models also capture information on case features. The legally fine-tuned BTO$_L$ performs slightly better than BTO$_M$ (MDR = 0.295, AUROC = 0.849 in column 13 versus MDR = 0.275, AUROC = 0.838 in column 10), suggesting that domain adaptation helps.

These results are relevant to legal debates on whether UDRP processes exhibit pro-complainant bias. While our correlative models *cannot* establish the absence of bias, our findings are consistent with the argument in [70,71] that high complaint success rates are better explained by case facts than structural pro-complainant biases. More importantly, our results suggest that the pipeline can automatically discover *correlative* legal predictors from decision texts. This becomes clearer when inspecting the discovered topics. The five LSA, BTO$_M$ and BTO$_L$ topics with smallest p-values in columns 7, 10 and 13, respectively, are presented in table 5. Some topics are intuitive. For example, the negative effect associated with LSA 17, a topic populated by n-gram variations on ‘administratively deficient’, suggests logically that ‘administratively deficient’ complaints correlate to worse complainant outcomes. Manual evaluation revealed that cases with the strongest weights for this topic indeed involved deficient complaints.[5] Three of these complaints were denied. Likewise, the top LSA three cases involved situations where the complainant provided incorrect ‘contact information’ for the domain registrant and was asked to amend the complaint accordingly.[6]

Other topics are less readable, but their underlying logic can be identified on closer inspection. For instance, LSA 19 and BTO$_M$ 5 are populated by references to famous trademarks and brands. These topics feature most strongly in complaints filed by large corporations which owned these and other famous marks, and typically against individuals who had registered variations on their brand names.[7] For example, 3 of the top 5 LSA 19 cases involved the ‘lego’ company suing for domains such as ‘legosets101.com’ and ‘legowolds.com’. Panels typically found evidence of bad faith in how respondents could not have registered these domains without knowing of the complainants’ well-known marks. Nine of the top 10 complaints succeeded. The exception was a complaint filed by ‘Hugo Boss’ for ‘boss-watch.com’ and ‘boss-world.com’. This was denied because the respondent had been selling watches under the ‘BOSS’ mark in Hong Kong since the 1970s, before the complainant’s mark was established.[8]

[5] WIPO Case nos. D2009-0021, D2011-1484, D2014-2277, D2014-1901 and D2016-0102.

[6] See WIPO Case nos. D2010-0593, D2012-1002, D2012-1761, D2014-1333 and D2016-0341.

[7] For LSA see WIPO Case nos. D2009-1392, D2013-1265, D2011-0391, D2010-1878, D2015-1445. For BTO$_M$ see WIPO Case nos. D2002-0760, D2006-0297, D2008-0416, D2011-0022, D2015-1936.

[8] WIPO Case no. D2015-1936.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 4. LASSO logit regression results for UDRP cases. Given the number of panellists and topics input we report medians and counts instead of individual estimates. Coefficients are direct estimates from `glmnet` and should not be interpreted cardinally. Standard errors in parentheses.

topic model:	none			OHE	LSA		
Y: complaint success (binary)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
panellists							
median coef	−0.0611	−0.0308	−0.0306	−0.037	−0.0397	−0.0222	−0.0241
median PV	0.0677	0.236	0.179	0.401	0.336	0.376	0.409
no. sig. ($\alpha = 0.05$)	53	50	49	19	9	13	8
no. sig. ($\alpha = 0.01$)	41	22	27	18	5	6	0
no. sig. ($\alpha = 0.001$)	12	8	19	1	1	1	0
no. selected	117	159	151	117	68	101	102
total no.	462	501	510	510	462	501	510
panel size		−0.154	−0.146	−0.0755		−0.138	−0.128*
		(0.024)	(0.023)	(0.023)		(0.0332)	(0.0236)
topics							
median coef				0.0374	0.0382	−0.0352	−0.0363
median PV				0.269	0.389	0.183	0.167
no. sig. ($\alpha = 0.05$)				7	24	31	32
no. sig. ($\alpha = 0.01$)				2	16	20	14
no. sig. ($\alpha = 0.001$)				1	9	8	6
no. selected				37	92	103	105
setting[a]	A	B	C	C	A	B	C
median dev. ratio	0.148	0.227	0.224	0.328	0.389	0.443	0.431
AUROC	0.742	0.824	0.822	0.874	0.903	0.914	0.914
	BTO_M			BTO_L			noise
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
panellists							
median coef	−0.0534	−0.0288	−0.0298	−0.0615	−0.0328	−0.0363	−0.0293
median PV	0.12	0.24	0.265	0.219	0.297	0.236	0.166
no. sig. ($\alpha = 0.05$)	39	40	47	25	42	8	52
no. sig. ($\alpha = 0.01$)	32	30	26	17	13	5	33
no. sig. ($\alpha = 0.001$)	22	16	9	9	2	4	24
no. selected	97	141	150	95	143	132	151
total no.	462	501	510	462	501	510	510
panel size		−0.185	−0.167		−0.125	−0.113	−0.143
		(0.0239)	(0.0233)		(0.0245)	(0.0234)	(0.023)

(*Continued.*)

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 4. (*Continued.*)

	BTO_M			BTO_L	noise		
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
topics							
median coef	−0.0267	0.027	−0.0114	0.032	0.0444	0.0295	0.0352
median PV	0.0854	0.351	0.317	0.206	0.654	0.348	0.58
no. sig. ($\alpha = 0.05$)	27	11	11	20	17	15	1
no. sig. ($\alpha = 0.01$)	15	4	4	10	10	9	0
no. sig. ($\alpha = 0.001$)	6	0	0	4	4	6	0
no. selected	69	97	96	67	79	85	30
setting[a]	A	B	C	A	B	C	C
median dev. ratio	0.222	0.274	0.274	0.215	0.29	0.295	0.237
AUROC	0.789	0.836	0.838	0.798	0.846	0.849	0.825

[a] Data partition and controls used. Setting A includes 20 150 1-panelist GTLD cases, excludes controls$_i$ (see §2c(i)), and sets $\lambda = 68.454$. B includes all 21 383 GTLD cases, includes controls$_i$ and sets $\lambda = 64.068$. C includes 22 653 GTLD/CCTLD cases, includes controls$_i$ and sets $\lambda = 61.784$.

Consider also BTO_L 69, which associates references to 'reverse domain name hijacking' (RDNH) with lower complaint success rates. UDRP Rule 1 defines RDNH as 'using the Policy in bad faith to attempt to deprive a registered domain name holder of a domain name'. When RDNH is found, the complaint fails. Recall however that the masked texts used in topic modelling exclude the 'Discussion and Findings' and 'Decision' sections, so the model should not have information on whether RDNH occurred. Inspecting the cases here reveals that RDNH n-grams feature strongly in the included 'Contentions' section when respondents actively defend the claim and raise the RDNH issue. In the usual case where respondents default, neither panellists nor complainants have incentives or need to discuss it. Thus while RDNH was not ultimately found in any of the top 5 BTO_L 69 cases, all were rare cases involving active respondents. This explains the topic's negative association with complaint success.

Not every topic can be easily understood. For instance, BTO_M 57, represented by n-grams referencing Middle Eastern countries, indeed involved complainants from this region.[9] Of these, two also involved Middle Eastern respondents. All five complaints were denied, but for differing reasons. In three cases the complainant failed to show bad faith because the domain had been registered *before* the complainant's mark was established. Whether complaints from Middle Eastern parties are properly associated with these facts and with lower success rates is however unclear. Likewise, BTO_L 1 is populated by n-grams tracking a typical portion in the 'Procedural History' section which states that 'the Panel has submitted the Statement of Acceptance and Declaration of Impartiality and Independence, as required by the Centre to ensure compliance with the Rules, paragraph 7'.[10] When this sentence occurs immediately before the next section header, 'Factual Background', the topic's n-grams arise (after stopword removal). Why this correlates with better complainant outcomes is not clear. It may signal the lack of other procedural issues, such that panellists can move directly to the next section,[11] but more qualitative evaluation is needed to ascertain this.

[9] WIPO Case nos. D2005-0309, D2008-0835, D2008-0895, D2009-0133 and D2015-0798.

[10] e.g. WIPO Case nos. D2006-0874, D2006-1054 and D2011-1122.

[11] e.g. WIPO Case no. D2010-0593.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 5. UDRP topics with smallest *p*-values across setting C regressions table 4: 7, 10 and 13. Coefficients are scaled estimates from the LASSO and should only be interpreted ordinally within the same regression. Topics are synthesized from masked decision texts that exclude 'Discussion and Findings' and later sections and should not be interpreted as capturing what the panels found.

model	topic	coefficient	representative n-grams
LSA	19	0.6069***	trade mark, lego, trade, famous, world, wipo case, amendment complaint, amendment, brand
	28	0.1724***	asserts, trade mark, complainant asserts, argues, complainant argues, alleges, complainant alleges, alleges respondent, trade
	3	0.3724***	contact information, registrant contact information, registrant contact, information, amended complaint, amended, amendment, amendment complaint, disclosed
	17	−0.4683***	trade mark, trade, amendment, deficient, administratively deficient, administratively, complaint administratively, complaint administratively deficient, amendment complaint
	2	−0.7685***	administrative, copy, e-mail, received, icann, notification, administrative panel, registrar domain, registrar domain name
BTO$_M$	5	0.1681**	armani, ikea, boss, bmw, reg, classes, elite, hugo, hugo boss, international trademark
	6	0.1517**	pharmaceutical, sanofi, pfizer, sanofiaventis, aventis, 100 countries, prescription, drug, treatment, weight
	11	0.1132**	chase, cme, barclays, financial services, bank, nasdaq, financial, insurance, investment, banking
	199	−0.044*	videos, sports, action, complaint exhibit, jeff, complaint exhibit respondent, january 31 2000, complainant action, skiing, omit
	57	−0.0747**	qatar, al, emirates, arabic, discover, project, uae, abu, brothers, dubai
BTO$_L$	1	0.2028***	rules paragraph factual, paragraph factual background, paragraph factual, factual background complainant, background complainant, panel submitted statement, ensure compliance rules, impartiality independence required, independence required, ensure compliance
	66	0.1911***	remedy transfer, support case, registered subsequently used, registered subsequently, elements complainant, subsequently used, complainant support, respondent transferred, remedy, policy domain
	21	0.1645***	publicdomainregistrycom, dba publicdomainregistrycom, pvt dba, directi internet solutions, directi internet, internet solutions, directi, internet solutions pvt, solutions pvt, pvt
	204	−0.0672***	oy, page displayed, banners, english version, portal, illegally, marks owned, domain names redirect, names redirect, alex
	69	−0.0754***	domain hijacking, reverse domain hijacking, hijacking, reverse domain, respondent requests, reverse, finding reverse, respondent requests panel, finding reverse domain, domain hijacking complainant

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 6. LASSO logit results for ECHR cases. λ's are separately derived per model following [77].

Y: violation found (binary)	Article 3			Article 6			Article 8		
topic model:	LSA	BTO$_M$	BTO$_L$	LSA	BTO$_M$	BTO$_L$	LSA	BTO$_M$	BTO$_L$
median coef	−0.164	−0.161	−0.144	0.153	−0.185	−0.119	−0.005	0.041	−0.129
median PV	0.02	0.177	0.088	0.521	0.171	0.324	0.211	0.257	0.451
no. sig. ($\alpha = 0.05$)	15	3	4	5	2	3	3	1	0
no. sig. ($\alpha = 0.01$)	13	1	4	3	1	1	2	0	0
no. sig. ($\alpha = 0.001$)	10	0	2	1	1	1	1	0	0
no. selected	28	13	9	17	5	21	10	4	16
N	568	568	568	903	916	916	457	458	458
λ	15.279	24.603	23.029	22.087	42.502	22.136	20.09	25.035	17.599
median dev. ratio	0.508	0.327	0.326	0.384	0.223	0.266	0.344	0.253	0.286
AUROC	0.89	0.73	0.739	0.855	0.746	0.784	0.763	0.638	0.711

(b) ECHR results

Table 6 presents results for LSA, BTO$_M$ and BTO$_L$ regressions fit on ECHR cases. As with the UDRP, LSA tends to produce higher model fits and the largest number of significant topics. This especially for Article 3, where 15 LSA topics are significant (at $\alpha = 0.05$) compared to 3 BTO$_M$ and 4 BTO$_L$ topics. This is notable given that LegalBERT was fine-tuned on ECHR cases [14]. It is thus not surprising that BTO$_L$ again produces higher fit measures than BTO$_M$, especially for Article 8 (AUROC = 0.711 versus 0.638). However, both BTO models produce broadly similar numbers of selected and significant topics.

Table 7 presents representative n-grams for significant ECHR topics chosen based on smallest p-value and largest coefficient sizes. For Article 3 (prohibition of torture or inhuman or degrading treatment), LSA 2 and BTO$_L$ 1 correctly discover and assign positive effects to what the ECtHR has described as 'a whole series of cases concerning allegations of disappearances in the Chechen Republic'.[12] Applicants were typically Chechen individuals whose close relatives were allegedly abducted by state military servicemen. Despite multiple complaints to and visits from the state's district prosecutor's office, the applicants hear nothing of their relatives for years. The ECtHR has 'found on many occasions' that the distress caused by their relatives' disappearance and the state's indifference to their plight violates Article 3. The top cases for these topics all involved similar fact patterns.[13]

BTO$_L$ 21 captures cases involving rejected asylum seekers who argued that they faced real risks of being subjected to treatment violating Article 3 if they were sent home. This allegedly because of their previous membership in military organizations that had clashed with their countries' current governments. 'December 2010' is a significant n-gram because the United Nations High Commissioner for Refugees had issued updated eligibility guidelines for Afghan asylum seekers then. The top 5 cases were all complaints from ex-Afghan security service personnel. As the negative coefficient suggests, these claims were typically denied because, among other reasons, these guidelines did not include them in their risk profiles for rights violations.[14] Notably, there is a similar line of unsuccessful complaints involving failed asylum

[12] HUDOC Case no. 001-140017.

[13] HUDOC Case nos. 001-140017, 001-112097, 001-95882, 001-95457, 001-92119, 001-93121, 001-150311, 001-146390 and 001-70853.

[14] HUDOC Case nos. 001-113328, 001-57451, 001-60924, 001-146372 and 001-69022.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Table 7. Significant ECHR topics across all Articles and topic models tested. We select topics to report here by first identifying the five lowest p-value topics within each regression, and then choosing topics with the three most positive and negative effects across all three regressions within each Article. Article 8 has only four significant topics in total.

Article	topic	coefficient	representative n-grams (by descending weight)
3	LSA 2	0.561***	servicemen, abduction, district prosecutor office, district prosecutor, prosecutor office, chechen, chechnya, military, prosecutor
	BTO_L 1	0.4994***	abduction, military, prosecutors office, men, relatives, prosecutors, criminal case, identify, forwarded, armed
	BTO_M 7	0.4069*	athe, bthe, governments, account, version, 1the, events, applicants detention, cthe, 2the
	BTO_L 21	−0.4696***	unhcr, december 2010, groups, sri, violations, 1951, lanka, sri lanka, international, refugees
	LSA 10	−0.7249***	regional, regional court, burned, villages, villagers, village, houses, pkk, appeal
	LSA 3	−1.3226***	asylum, country, refugee, board, unhcr, kingdom, forces, united, members
6	LSA 2	1.5781***	court, applicant, enforcement, russian, bailiffs, ukraine, ukrainian, uah, enforcement proceedings
	LSA 6	0.3544*	ukraine, uah, ukrainian, bailiffs, bailiffs service, debtor, turkish, v., law relevant
	LSA 11	0.2906**	applicants, tenant, land, italian, cell, property, detention, possession, treatment
	BTO_L 79	−0.2929*	greek, territory, entry, called, witnesses, level0, level0 arabic, seq level0, seq level0 arabic, arabic
	BTO_L 0	−0.3873***	3the applicant, alleged, 3the, detention, access, complained, custody, statements, particular, torture
	BTO_M 0	−0.5878***	imprisonment, years, expert, village, tax, damage, social, seq, medical, regional court
8	LSA 2	0.822***	detention, prosecutor, criminal, remand, applicant detention, criminal proceedings, regional, regional court, applicant
	LSA 10	0.3846**	hague convention, russian, poland, gypsies, polish, gypsy, hague, retention, sites
	BTO_M 70	0.3753*	correspondence, prohibition, mean, content, shown, february 2003, ask, world, avoid, 9the
	LSA 7	−0.3084*	prison, detention, expulsion, aliens, residence permit, detained, visits, asylum, cell

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

seekers who previously served in the Sri Lankan Tamil Tigers.[15] These were also picked up by LSA 12 (not tabulated in table 7), which is represented by n-grams including 'sri lanka', 'ltte' and 'colombo'. LSA 3 also identifies cases involving unsuccessful asylum seekers, but includes more varied claims from individuals originally from Somalia, Iraq and Libya.[16] These complaints tended to fail because the court did not find a sufficiently real risk of treatment contrary to Article 3.

For Article 6 (right to fair trial), LSA 2 and 6 surface a collection of cases where Ukrainian individuals awarded compensation judgments against certain (often state-linked) companies were forced to wait for years before receiving due payment. They argued that the state bailiff had inordinately delayed enforcement proceedings. Decisions for such cases are worded very similarly, and typically reiterate how the ECtHR has 'already' or 'frequently found' violations in like cases.[17] Notably, a Ukrainian government judicial enforcement reform effort acknowledges

[15] HUDOC Case nos. 001-102949, 001-102955, 001-102947, 001-102957 and 001-104956.

[16] HUDOC Case nos. 001-145018, 001-118339, 001-145789, 001-126027 and 001-141949.

[17] HUDOC Case nos. 001-91393, 001-71592, 001-93886, 001-75842, 001-78383, 001-78528, 001-78530, 001-75842, 001-78397 and 001-70357.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

Article 6 as its motivation.[18] BTO$_L$ 79 is diluted by markup n-grams like 'level0' but nonetheless identifies several cases involving Cypriot individuals who had participated in a 1989 anti-Turkish demonstration in disputed territory arising out of the 1974 Turkish intervention in North Cyprus.[19] They were charged and convicted in the Turkish courts for entering Turkish territory without permission. Typically, they argued that their Article 6 rights had been violated because the legal proceedings were generally in Turkish, not Greek which they understood. The ECtHR generally rejected these claims because the applicants had reasonable access to interpreters and other legal assistance.

For Article 8 (right to respect for family and private life, home and correspondence), LSA 2 points to a line of cases filed against the Polish authorities by individuals detained in criminal remand over the authorities' standard practice of reading and re-sealing correspondence sent by these individuals to the courts and stamping the envelopes with a 'censored' label. The ECtHR has noted how it has 'held on many occasions' that the label forces the court to assume an interference with correspondence that, unless justified, violates Article 8.[20] Two of the top 5 BTO$_M$ 70 cases have similar facts, but the topic also seems to cover other kinds of interferences to correspondence. Relatedly, the top 5 LSA 10 cases all involve complaints filed by Romani persons[21] against the British government and its consistent refusal to grant them planning permission to develop land they owned into caravan sites. After the ECtHR found this to be a violation in 1996,[22] several similar and ultimately successful cases were raised in which the ECtHR expressly 'recalls that it has already examined' such complaints and found violations.[23]

4. Discussion

We proposed and evaluated an automated pipeline for discovering significant topics from cases by performing penalized regression and selective inference on features synthesized from decision texts using topic models. We show that significant topics discovered through this process capture relevant information on factual patterns correlated with case outcomes. On a large (by legal standards) dataset of UDRP cases, legal outcome models fitted with decision text topics consistently produce higher fit scores compared with models fitted without (table 4). The LASSO also tends to select the topics as significant predictors over other structured case attributes of potential interest, such as judge identities. Coefficients and p-value estimates also change noticeably. This holds across several regression settings and topic modelling approaches. Using a canonical dataset of ECHR cases, we show that the method generalizes relatively easily, without the need for additional feature engineering or pre-processing. Only structured outcome information and unstructured decision texts are required, though additional variables can be added at the regression step. Running similar procedures on the existing dataset, albeit with corpus-tailored hyperparameters (such as topic number and λ), yields significant topics consistent with ECHR case law.

Our experiments show that LSA is a useful, even if dated, codebook for decision texts. Across all experiment settings, LSA produced higher fit scores and a higher number of significant topics than both BTO models. LSA is also computationally cheaper. This may appear counterintuitive since BTO is a significantly more sophisticated model which exploits recent advances like word embeddings and language modelling. As noted by Soh *et al.* [80] in the context of legal topic classification, the length of legal decision texts may offer one explanation for LSA out-

[18] www.kmu.gov.ua/en/reformi/verhovenstvo-prava-ta-borotba-z-korupciyeyu/reformuvannya-sistemi-vikonannya-sudovih-rishen.

[19] HUDOC Case nos. 001-169203, 001-61582, 001-139903, 001-113876 and 001-139995.

[20] HUDOC Case no. 001-93604 at paragraph 94.

[21] We have preferred the term 'Romani' throughout the article as the token 'gypsy' has derogatory connotations. With apologies to Romani people, the original token was retained in table 7 to reflect what was actually used in the corpus.

[22] HUDOC Case no. 001-58076.

[23] HUDOC Case nos. 001-59156, 001-59154, 001-59158 and 001-59157.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

performing newer approaches in the legal domain. BTO's superiority over traditional topic models has mainly been demonstrated on shorter texts like tweets and news articles [26,81]. For longer documents, our present approach of reconstituting document-level topics by normalizing chunk-level topic counts is, while standard in the literature [40], unlikely to be the optimal way to deploy BTO. Using LMs with larger context windows than those tested here could significantly improve BTO's performance. Thus, we do not suggest that LSA is necessarily better-suited for this task. Further, since each topic model yields different topics which may provide different insights on the cases, there is no clear metric for 'better' in this context. Additionally, BTO's lower fit scores may be a methodological artefact since we did not conduct hyperparameter optimization, but chose similar parameters across all topic models to establish a baseline comparison. A fully optimized BTO model may outperform a fully optimized LSA. Hyperparameter tuning was not done because, unlike typical machine learning settings, our task focuses on explanation rather than classification and does not yield any clear performance metric (e.g. F1 score) for evaluating a grid search. BTO's performance here should be interpreted in this light.

More importantly, qualitative evaluation of significant topic n-grams demonstrates that the discovered topics rest on sound and interpretable legal bases. For UDRP cases, the topics shed correlative light on how administratively deficient complaints are less likely to win, how famous trademark owners can be associated with higher success rates, and how respondents who actively defend their domains are less likely to lose. For Articles 3, 6 and 8 ECHR, the topics identify archetypal cases involving abducted Chechen relatives, Afghan asylum seekers, Ukrainian judgment enforcers, Cypriot demonstrators, Polish detainees, and Romani land owners. These are correctly associated with their usual case outcomes. Essentially, unique case features prompt judges into writing decisions with a higher preponderance of correspondingly unique n-grams, producing signals which the topic models are capable of detecting. As legal decisions are written with close reference to case facts and relevant laws, and judges would generally not write about irrelevant matters and non-issues, we theorize that the decision text generation function accords with the standard topic modelling assumption that texts are generated by sampling n-grams from latent topics [82].

To be sure, not every discovered topic made sense. This may point to limitations in our evaluation process, since we only sampled the top five cases associated with the most significant topics. We may also have been unable to detect known patterns that the topics were in fact referencing. Further, there is no reason why each topic should capture exactly one predictor. Certain topics may have had n-gram distributions amalgamating several archetypal case features. Individually insignificant topics could have been jointly significant with others. Future work could examine this further by modelling interaction terms and conducting joint tests, though this may make interpreting the topics and coefficients more challenging for evaluators. Besides human limitations, the automated process is also imperfect. It can produce false positives (e.g. significant topics which do not actually capture legal predictors; high document-topic weights for a case not actually on topic) and false negatives (not attaching significance to topics which do; not synthesizing a related topic to begin with).

At a more abstract level, the four pipeline steps can be understood as a series of dimensionality reduction steps, starting with a large decision corpus, and resulting in numerical associations between decisions and topics (document-topic weights), topics and outcomes (regression coefficients) and topics and words (n-gram distributions). These mappings can be analysed transitively to assist with the discovery, extraction, and analysis tasks identified in §2(a). To illustrate, after observing UDRP LSA topic 19 (table 5), researchers could create and extract observations for an indicator variable for whether the complainant owned a famous mark. Decision-topic weights produced by the method could guide the extraction process. After several such variables are extracted, a regression (not necessarily the LASSO) could then be run on the reduced dataset. Notably, given the increasing popularity of large language models, the pipeline's ability to reduce large legal corpora to smaller components could prove useful in fitting legal texts into limited context windows.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

5. Conclusion

We proposed and assessed an automated method for discovering significant topics given only decision texts and case outcomes, building on prior work examining how topic models can be used to predict and explain case outcomes [7,8,11]. The task of legal topic discovery was formally defined and distinguished from related identification and analysis tasks. We developed and demonstrated pre-processing, topic modelling, regression and inference steps tailored to this task and its legal context. The method shows promise in its ability to discover archetypal case features and patterns consistent with the jurisprudence of the UDRP and ECHR datasets tested, and could generalize to other areas. It is however not perfect, and should be applied bearing the possibility of false positives and negatives in mind. There are two extensions we hope to pursue in future work. First, to conduct more rigorous experiments and hyperparameter search with BERTopic and its variations. Notably, BERTopic is a modular framework involving six steps that accept several different algorithms and (optional) parameters. Second, a more robust yet ideally less manual method for evaluating and interpreting topics could be developed.

Data accessibility. This article has no additional data.

Declaration of AI use. I have not used AI-assisted technologies in creating this article.

Author contributions. J.S.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—review and editing.

Conflict of interest declaration. I declare I have no competing interests.

Funding. No funding has been received for this article.

Acknowledgements. I thank Georgios Georgiou, James Greiner, Alma Cohen, Lucian Bebchuk, William Hubbard, Michael Livermore and Kevin Ashley, and participants of the Complexity in Law & Governance Works in Progress Workshop for constructive comments, including on very early versions of this work. I thank also the editors of this special collection Daniel Katz, JB Ruhl, and Pierpaolo Vivo, as well as the dedicated editorial staff at PhilTrans A. All errors are mine.

References

1. McCarty LT. 2007 Deep semantic interpretations of legal texts. In *Proc. Int. Conf. on Artificial Intelligence and Law, Stanford, CA, USA, 4–8 June 2007*, pp. 217–224. New York, NY: Association of Computing Machinery.
2. Nazarenko A, Lévy F, Wyner A. 2021 A pragmatic approach to semantic annotation for search of legal texts: an experiment on GDPR. In *Proc. Int. Conf. on Legal Knowledge and Information Systems, Vilnius, Lithuania, 8–10 December 2021*, pp. 23–32. Amsterdam, The Netherlands: IOS Press.
3. Grimmer J, Stewart BM. 2013 Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Pol. Anal.* **21**, 267–297. (doi:10.1093/pan/mps028)
4. Alschner W, Pauwelyn J, Puig S. 2017 The data-driven future of international economic law. *J. Int. Econ. Law* **20**, 217–231. (doi:10.1093/jiel/jgx020)
5. Choi J. 2020 An empirical study of statutory interpretation in tax law. *N. York Univ. Law Rev.* **95**, 363–441. (doi:10.2139/ssrn.3460962)
6. Soh J. 2019 A network analysis of the Singapore Court of Appeal's citations to precedent. *Singapore Acad. Law J.* **31**, 246–284. (doi:10.2139/ssrn.3346422)
7. Aletras N, Tsarapatsanis D, Preoţuc-Pietro D, Lampos V. 2016 Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ Comput. Sci.* **2**, e93. (doi:10.7717/peerj-cs.93)
8. Falakmasir MH, Ashley KD. 2017 Utilizing vector space models for identifying legal factors from text. In *Proc. Int. Conf. on Legal Knowledge and Information Systems, Luxembourg, 13–15 December 2017*, pp. 183–192. Amsterdam, The Netherlands: IOS Press.
9. Carter DJ, Brown JJ, Rahmani A. 2016 Reading the High Court at a distance: topic modelling the legal subject matter and judicial activity of the High Court of Australia, 1903–2015. *Univ. New South Wales Law J.* **39**, 1300–1354. (doi:10.31228/osf.io/qhezc)
10. Dyevre A, Glavina M, Ovádek M. 2021 The voices of European law: legislators, judges and law professors. *German Law J.* **22**, 956–982. (doi:10.1017/glj.2021.47)

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

11. Salaün O, Gotti F, Langlais P, Benyekhlef K. 2022 Why do tenants sue their landlords? Answers from a topic model. In *Proc. Int. Conf. on Legal Knowledge and Information Systems, Saarbrücken, Germany, 14–16 December 2022*, pp. 113–122. Amsterdam, The Netherland: IOS Press.
12. Dhanani J, Mehta R, Rana D. 2022 Effective and scalable legal judgment recommendation using pre-learned word embedding. *Complex Intell. Syst.* **8**, 3199–3213. (doi:10.1007/s40747-022-00673-1)
13. Jayasinghe S, Rambukkanage L, Silva A, de Silva N, Perera S, Perera M. 2022 Learning sentence embeddings in the legal domain with low resource settings. In *Proc. 36th Pacific Asia Conf. on Language, Information and Computation, Manila, The Philippines, 20–22 October 2022*, pp. 494–502. Stroudsburg, PA: Association for Computational Linguistics.
14. Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. 2020 LEGAL-BERT: the muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020, 16–20 November 2020*, pp. 2898–2904. Stroudsburg, PA: Association for Computational Linguistics.
15. Paul S, Mandal A, Goyal P, Ghosh S. 2023 Pre-trained language models for the legal domain: a case study on Indian law. In *Proc. Int. Conf. on Artificial Intelligence and Law, Braga, Portugal, 19–23 June 2023*, pp. 187–196. New York, NY: Association of Computing Machinery.
16. Licari D, Bushipaka P, Marino G, Comandé G, Cucinotta T. 2023 Legal holding extraction from Italian case documents using Italian LEGAL-BERT text summarization. In *Proc. Int. Conf. on Artificial Intelligence and Law, Braga, Portugal, 19–23 June 2023*, pp. 148–156. New York, NY: Association of Computing Machinery.
17. Liu Z, Chen H. 2017 A predictive performance comparison of machine learning models for judicial cases. In *IEEE Symp. Series on Computational Intelligence, Honolulu, HI, USA, 27 November–1 December 2017*, pp. 1–6. New York, NY: Institute of Electrical and Electronics Engineers.
18. Chalkidis I, Androutsopoulos I, Aletras N. 2019 Neural legal judgment prediction in English. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019*, pp. 4317–4323. Stroudsburg, PA: Association for Computational Linguistics.
19. Medvedeva M, Wieling M, Vols M. 2023 Rethinking the field of automatic prediction of court decisions. *Artif. Intell. Law* **31**, 195–212. (doi:10.1007/s10506-021-09306-3)
20. Medvedeva M, Vols M, Wieling M. 2020 Using machine learning to predict decisions of the European Court of Human Rights. *Artif. Intell. Law* **28**, 237–266. (doi:10.1007/s10506-019-09255-y)
21. Ashley KD, Brüninghaus S. 2009 Predicting outcomes of case based legal arguments. *Artif. Intell. Law* **17**, 125–165. (doi:10.1007/s10506-009-9077-9)
22. Branting K, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, Pfaff M, Liao B. 2021 Scalable and explainable legal prediction. *Artif. Intell. Law* **29**, 213–238. (doi:10.1007/s10506-020-09273-1)
23. Gray M, Savelka J, Oliver W, Ashley K. 2022 Toward automatically identifying legally relevant factors. In *Proc. Int. Conf. on Legal Knowledge and Information Systems, Saarbrücken, Germany, 14–16 December 2022*, pp. 53–62. Amsterdam, The Netherlands: IOS Press.
24. Gray MA, Savelka J, Oliver W, Ashley K. 2023 Automatic identification and empirical analysis of legally relevant factors. In *Proc. Int. Conf. on Artificial Intelligence and Law, Braga, Portugal, 19–23 June 2023*, pp. 101–110. New York, NY: Association of Computing Machinery.
25. Landauer TK, Foltz PW, Laham D. 1998 Introduction to latent semantic analysis. *Discourse Process.* **25**, 259–284. (doi:10.1080/01638539809545028)
26. Grootendorst M. 2022 BERTopic: neural topic modeling with a class-based TF-IDF procedure.
27. Sulea OM, Zampieri M, Vela M, Genabith JV. 2017 Predicting the law area and decisions of French supreme court cases. In *Proc. Int. Conf. on Recent Advances in Natural Language Processing, Varna, Bulgaria, 2–8 September 2017*, pp. 716–722. Stroudsburg, PA: Association for Computational Linguistics.
28. Tibshirani R. 1994 Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288. (doi:10.1111/j.2517-6161.1996.tb02080.x)
29. Taylor J, Tibshirani R. 2017 Post-selection inference for 1-penalized likelihood models. *Can. J. Stat.* **46**, 41–61. (doi:10.1002/cjs.11313)

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

30. Canavotto I, Horty J. 2023 Reasoning with hierarchies of open-textured predicates. In *Proc. Int. Conf. on Artificial Intelligence and Law, Braga, Portugal, 19–23 June 2023*, pp. 52–61. New York, NY: Association of Computing Machinery.
31. Chen DL. 2023. Judicial analytics and the great transformation of American law. *Artif. Intell. Law* **27**, 15–42. (doi:10.1007/s10506-018-9237-x)
32. Ruger TW, Kim PT, Martin AD, Quinn KM. 2004 The Supreme Court forecasting project: legal and political science approaches to predicting supreme court decisionmaking. *Columbia Law Rev.* **104**, 1150. (doi:10.2307/4099370)
33. Soh J. 2021 Causal inference with legal texts. *MIT Computational Law Report*. See https://law.mit.edu/pub/causalinferencewithlegaltexts.
34. Cinelli C, Forney A, Pearl J. 2022 A crash course in good and bad controls. *Sociol. Methods Res.* (doi:10.1177/00491241221099552)
35. Rowland B. 2014 Writing a statement of facts in an appellate brief. See https://www.law.georgetown.edu/wp-content/uploads/2018/07/StatementofFactsinaBriefFinal.pdf (accessed 30 June 2023).
36. Jones KS. 1972 A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **28**, 11–21. (doi:10.1108/eb026526)
37. Salton G, Buckley C. 1988 Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* **24**, 513–523. (doi:10.1016/0306-4573(88)90021-0)
38. Dominich S. 2001 *Mathematical foundations of information retrieval*. Dordrecht, The Netherlands: Springer Science Business Media.
39. Bishop C. 2006 *Pattern recognition and machine learning*. Berlin, Germany: Springer.
40. Silveira R, Fernandes CGO, Neto JAM, Furtado V, Filho JEP. 2021 Topic modelling of legal documents via LEGAL-BERT. In *Proc. Int. Workshop on Relations in the Legal Domain, Sao Paulo, Brazil, 25 June 2021*, pp. 64–72.
41. McInnes L, Healy J, Saul N, Großberger L. 2018 UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861. (doi:10.21105/joss.00861)
42. McInnes L, Healy J, Astels S. 2017 hdbscan: hierarchical density based clustering. *J. Open Source Softw.* **2**, joss00205. (doi:10.21105/joss.00205)
43. HuggingFace. 2021 all-MiniLM-L6-v2. See https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
44. Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. 2020 MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proc. Int. Conf. on Advances in Neural Information Processing Systems, Vancouver, Canada, 6–12 December 2020*, pp. 5776–5788. New York, NY: Curran Associates.
45. Devlin J, Chang MW, Lee K, Toutanova K. 2019 BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019*, pp. 4171–4186. Stroudsburg, PA: Association for Computational Linguistics.
46. Scikit-Learn. 2017 Truncated singular value decomposition and latent semantic analysis. See http://scikit-learn.org/stable/modules/decomposition.html#lsa (accessed 30 June 2023).
47. Zou H, Hastie T. 2005 Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67**, 301–320. (doi:10.1111/j.1467-9868.2005.00503.x)
48. Hastie T, Tibshirani R, Friedman JH. 2017 *The elements of statistical learning: data mining, inference, and prediction*. Berlin, Germany: Springer.
49. Zorn C. 2005 A solution to separation in binary response models. *Pol. Anal.* **13**, 157–170. (doi:10.1093/pan/mpi009)
50. Bielza C, Robles V, Larrañaga P. 2011 Regularized logistic regression without a penalty term: an application to cancer classification with microarray data. *Expert Syst. Appl.* **38**, 5110–5118. (doi:10.1016/j.eswa.2010.09.140)
51. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. 2015 Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat. Med.* **35**, 1159–1177. (doi:10.1002/sim.6782)
52. Firth D. 1993 Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38. (doi:10.1093/biomet/80.1.27)

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024

53. Heinze G. 1999 *The application of Firth's procedure to Cox and logistic regression*. Technical Report 10. Vienna, Austria: Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna.
54. Hastie T, Tibshirani R, Wainwright M. 2015 *Statistical learning with sparsity: the lasso and generalizations*. Boca Raton, FL: CRC Press.
55. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009 Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721. (doi:10.1093/bioinformatics/btp041)
56. Zhu Y, Tan TL, Cheang WK. 2017 Penalized logistic regression for classification and feature selection with its application to detection of two official species of ganoderma. *Chemom. Intell. Lab. Syst.* **171**, 55–64. (doi:10.1016/j.chemolab.2017.09.019)
57. Lee JD, Sun DL, Sun Y, Taylor JE. 2016 Exact post-selection inference, with application to the lasso. *Ann. Stat.* **44**, 907–927. (doi:10.1214/15-aos1371)
58. Tibshirani R, Tibshirani R, Taylor J, Loftus J, Reid S. 2017 selectiveInference: tools for post-selection inference. R package version 1.2.4.
59. Friedman J, Hastie T, Tibshirani R. 2010 Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22. (doi:10.18637/jss.v033.i01)
60. Katz DM, Bommarito MJ, Blackman J. 2017 A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE* **12**, e0174698. (doi:10.1371/journal.pone.0174698)
61. Chik WB. 2007 Lord of your domain, but master of none: the need to harmonize and recalibrate the domain name regime of ownership and control. *Int. J. Law Inf. Technol.* **16**, 8–72. (doi:10.1093/ijlit/eam005)
62. Mueller M. 2002 *Ruling the root: internet governance and the taming of cyberspace*. New York, NY: MIT Press.
63. Epstein L, Weinshall K. 2021 *The strategic analysis of judicial behavior*. Cambridge, UK: Cambridge University Press.
64. Brannigan C. 2004 The UDRP: how do you spell success? *Digital Technol. Law J.* **5**, 2.
65. Mueller M. 2001 Rough justice: a statistical assessment of ICANN's uniform dispute resolution policy. *Inf. Soc.* **17**, 151–163. (doi:10.1080/01972240152493029)
66. Geist MA. 2002 Fair.com? An examination of the allegations of systemic unfairness in the ICANN UDRP. *Brookings J. Int. Law* **27**, 903. (doi:10.2139/ssrn.280630)
67. Mueller M. 2002 *Success by default: a new profile of domain name trademark disputes under ICANN's UDRP*. Syracuse University School of Information Studies.
68. Kelley PD. 2002 Emerging patterns in arbitration under the uniform domain-name dispute-resolution policy. *Berkeley Technol. Law J.* **17**, 181–204. (doi:10.15779/Z38K39R)
69. Donahey MS. 2001 The UDRP: fundamentally fair, but far from perfect. *Electron. Commerce Law Rep.* **6**.
70. Ned B. 2002 *UDRP—a success story: a rebuttal to the analysis and conclusions of Professor Milton Mueller in 'Rough Justice'*. New York, NY: International Trademark Association.
71. Kur A. 2002 *UDRP*. Max-Planck-Institute for Foreign and International Patent, Copyright and Competition Law, Munich.
72. Kesan JP, Gallo AA. 2005 The market for private dispute resolution services: an empirical re-assessment of ICANN-UDRP performance. *Michigan Telecommun. Technol. Law Rev.* **11**, 285–380. (doi:10.2139/ssrn.688001)
73. Klerman D. 2017 Forum selling and domain-name disputes. *Loyola Univ. Chicago Law J.* **48**, 561–584.
74. Greene D, O'Callaghan D, Cunningham P. 2014 How many topics? Stability analysis for topic models. In *ECML/PKDD*. Lecture Notes in Computer Science, vol. 8724, pp. 498–513. Berlin, Germany: Springer.
75. Gerlach M, Peixoto TP, Altmann EG. 2018 A network approach to topic models. *Sci. Adv.* **4**, eaaq1360. (doi:10.1126/sciadv.aaq1360)
76. Sbalchiero S, Eder M. 2020 Topic modeling, long texts and the best number of topics. Some problems and solutions. *Qual. Quant.* **54**, 1095–1108. (doi:10.1007/s11135-020-00976-w)
77. Negahban S, Yu B, Wainwright MJ, Ravikumar PK. 2009 A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Proc. Int. Conf. on Advances in Neural Information Processing Systems, Vancouver, Canada, 10–16 December 2009*, pp. 1348–1356. New York, NY: Curran Associates.

78. McShane BB, Wyner AJ. 2011 A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable? *Ann. Appl. Stat.* **5**, 5–44. (doi:10.1214/10-AOAS398)
79. Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz D, Aletras N. 2022 LexGLUE: a benchmark dataset for legal language understanding in English. In *Proc. Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022*, pp. 4310–4330. Stroudsburg, PA: Association for Computational Linguistics.
80. Soh J, Lim HK, Chai IE. 2019 Legal area classification: a comparative study of text classifiers on Singapore Supreme Court judgments. In *Proc. Natural Legal Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019*, pp. 67–77. Stroudsburg, PA: Association for Computational Linguistics.
81. de Groot M, Aliannejadi M, Haas MR. 2022 Experiments on generalizability of BERTopic on multi-domain short text. In *Proc. Widening NLP Workshop, Abu Dhabi, UAE, 7 December 2022*, p. 70. Stroudsburg, PA: Association of Computational Linguistics.
82. Blei DM, Ng AY, Jordan MI. 2003 Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022.

Downloaded from https://royalsocietypublishing.org/ on 27 March 2024