Return of the AI: An analysis of legal research on Artificial Intelligence using topic modeling

Return of the AI: An Analysis of Legal Research on Artificial Intelligence Using Topic Modeling

Constanta Rosca, Bogdan Covrig, Catalina Goanta, Gijs van Dijck, Gerasimos Spanakis
{constanta.rosca,b.covrig,catalina.goanta,gijs.vandijck,jerry.spanakis}@maastrichtuniversity.nl
Law & Tech Lab, Maastricht University
Maastricht, Netherlands

ABSTRACT

AI research finds itself in the third boom of its history, and in recent years, AI-related themes have gained considerable popularity in new disciplines, such as law. This paper explores what legal research on AI constitutes of and how it has evolved, while addressing the issues of information retrieval and research duplication. Using Latent Dirichlet Allocation (LDA) topic modeling on a dataset of 3931 journal articles, we explore three questions: (a) Which topics within legal research on AI can be distinguished? (b) When were these topics addressed? and (c) Can similar papers be detected? The topic modeling results in a total of 32 meaningful topics. Additionally, it is found that legal research on AI drastically increased as of 2016, with topics becoming more granular and diverse over time. Finally, a comparison of the similarity assessments produced by the algorithm and a human expert suggest that the assessments often coincide. The results provide insights into how a legal research on AI has evolved over time, and support for the development of machine learning and information retrieval tools like LDA that assist in structuring large document collections and identifying relevant articles.

CCS CONCEPTS

• **Computing methodologies** → **Topic modeling**; • **Applied computing** → **Law**.

KEYWORDS

information retrieval, topic modeling, legal research

ACM Reference Format:
Constanta Rosca, Bogdan Covrig, Catalina Goanta, Gijs van Dijck, Gerasimos Spanakis. 2020. Return of the AI: An Analysis of Legal Research on Artificial Intelligence Using Topic Modeling. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Artificial intelligence (AI) research finds itself in the third boom of its history, fuelled by increased funding, scientific breakthroughs such as deep learning, and widespread public speculation relating to the scope and impact of these breakthroughs. While decades ago the interest in AI research was generally limited to specific disciplines (e.g. computer science, philosophy), during the past years AI-related themes have gained considerable popularity in new disciplines, such as law.

With over 2500 publications already by the year 2015 referring to 'artificial intelligence' [19] (see also Figure 3 in the Appendix), it may no longer be realistic to assume that researchers can keep up with legal research on AI, or the number of publications in general. Moreover, the recent spike suggests more and more authors have started writing about AI in law topics, including authors who have previously not published on such topics. This, in combination with the inability to keep up with legal research on AI due to its exponential growth, creates the risk that authors replicate previous work without being aware of similar previous publications.

This paper aims to explore what legal research on AI constitutes of and how it has evolved while addressing the issue of information retrieval and the risk of research duplication. We develop a methodology that distinguishes topics in a collection of documents (in this case journal publications), allows exploring the evolution of the topics over time, and detects similarity between documents, with the purpose of providing solutions for reading and analyzing a number of publications in bulk in ways that humans cannot. Consequently, we aim to answer the following research questions (RQ):

RQ1: Which topics within the field of legal research on AI can be distinguished ('What')? A methodology would provide insight in how legal research on AI is structured, and it would allow classifying publications in sub-topics, which would enhance information retrieval.

RQ2: What (i.e. about which topics) has been written when ('What - When')? This question contributes to the understanding of which topics have emerged, remained, or lost the interest of legal scholars. The analysis of the What - When question will provide information about the evolution of legal research on AI.

RQ3: Can similar papers be detected? Considering the sharp increase of publications, it may be becoming increasingly difficult to find publications on similar research questions, which may even result in reproduction of scholarship because prior publications are overlooked. This question explores a methodology that allows detecting thematically similar documents in a given corpus.

2 BACKGROUND

One of the innovations in this paper is to use unsupervised machine learning to categorize legal research on AI and to map how legal scholarship on this topic has developed. The idea of smart literature reviews has received prior attention, on the grounds of how manual searches for existing literature - especially in matured domains where scholarship is abundant - is not efficient and might

NLLP @ KDD 2020, August 24th, San Diego, US

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

have a negative impact on the quality of new research [2]. Similar approaches have been taken to map computer science literature [23], as well as communications research [29].

To the best of our knowledge, legal scholarship as such has not yet been analyzed using LDA. Within the narrow confines of the field of research labeled as AI and Law, a lot of legal research on information retrieval has been published in the past decades. Notwithstanding that the volume of scholarship on information retrieval pertaining to the discipline of computer science alone is vast to say the least [13], particularly when applied to the legal field, such research focused on the availability of legal information [5, 20, 21, 28, 32, 34, 46, 55], search systems and search strategies [16, 22, 30, 43, 47, 57], information processing [6, 17, 18, 25, 33, 36, 37, 50], and the role of legal publishers [1, 3, 26]. These publications address a variety of issues and questions, including the sustainability of publicly available legal repositories (e.g. AustLII), the importance of natural language processing when searching for legal information, the performance of online searches compared to searching through paper, how citation analysis may be used to improve search results, the role of legal publishers in this and, more generally, the impact of automation on how the law is analyzed and applied.

Still, the question of how to capture and visualize the development of an entire legal sub-field like legal research on AI has barely been explored. One of the avenues of exploration has been shaped by Bench-Capon et al., who have previously focused on the proceedings of the International Conference on AI and Law, first held in 1987, to make a 25-year retrospective of the research generated therein by describing the scholarship progress through illustrative papers selected from various editions [4]. Other studies focused on traditional (systematic) literature reviews on the impact of AI on specific legal domains, such as administrative law [40] or intellectual property [24].

However, given the limitations of legal databases or the way they are used by researchers, making comprehensive overviews of existing literature remains a considerable hurdle. This is all the more so in the past four years, when the production of legal scholarship on artificial intelligence seems to have grown considerably (see Figure 3 in the Appendix). This paper aims to fill this research gap by proposing and testing an unsupervised machine learning approach to the clustering of literature in this field, namely topic modeling.

3 METHODOLOGY

3.1 Corpus

The corpus includes a total of 3931 journal articles obtained from the HeinOnline database. Absent a centralised, comprehensive, open access repository for international legal scholarship, we focused on one of the commercially available databases. HeinOnline is one of the leading international databases on legal materials, which contains over 170 million pages of literature and indexing over 2700 law journals. In the section 'Law Journal Library', section type 'Articles', the corpus covers literature available in the database between 1960 and 2018. Unlike arXiv, HeinOnline does not have a section on 'artificial intelligence'. The total number of retrieved articles reflects the results of a boolean search using the keywords 'artificial intelligence', namely all articles which include both terms. Resulting articles therefore discuss a wide array of aspects relating to artificial intelligence, and do not, as such, focus on specific technical or legal issues. For the purpose of this study, we assume that even one reference to the keywords is sufficient to include an article in the corpus.

The articles in the corpus follow a power law distribution, where a relatively large number of publications is written by a low number of authors, and few publications by many authors (see Figure 4 in the Appendix). The same distribution applies to the number of publications per author(s) with roughly 28 authors having more than 5 publications (see Figure 5 in the Appendix).

3.2 Topic Modeling

3.2.1 Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) [8] was used to identify topics in the corpus. LDA has many different use cases so far. Examples concern organizing large document collections in order to improve search and retrieval of information, summarization of large textual data, and even image clustering. In the legal domain, LDA has been used to study the agenda of the US Supreme Court [27], the High Court of Australia [11] and the Court of Justice of the European Union (CJEU) [15]. Winkels [58] used LDA to build a recommender system for Dutch case law. Panagis and Sadl [39] combined network analysis and LDA to study the case-law generative process of the CJEU.

LDA is a generative, probabilistic model for a collection of documents, which are represented as mixtures of latent topics, where each topic is characterized by a distribution over all words of the collection of documents. The basic representation unit of the documents is the word, i.e. all distinct terms are extracted from the document collection along with their frequencies (per document), which is the so called 'Bag-of-Words' model [41].

On a conceptual level, the algorithm tries to discover *topics* that can represent the collection of documents. Each document is generated from a mixture of these topics and each topic is generated from a probability vector (distribution) over all words. Assuming such a generative model for any collection of documents, LDA's goal is to try and 'backtrack' this process, i.e. find a set of topics that are likely to have generated the whole document collection.

3.2.2 LDA Variants. In this paper, we used the model of Blei et al. [8], however we did explore other variations as well. Stevens et al. present a qualitative comparison of different topic model algorithms, [48] also using different evaluation metrics [31, 35] and conclude that, given the same data and the same number of topics, LDA is able to learn more coherent topics than the competing approaches.

Nevertheless, there are extensions of the LDA model towards topic tracking over time [53, 56]. However, according to Wang et al. [52], these methods deal with constant topics and the timestamps are used for better discovery. Opposed to that, in a Blei et al. paper [7] a model for detection of evolving topics in a discrete time space is presented (Dynamic Topic Modeling - DTM). Here, LDA is used on topics aggregated in time epochs and a state space model handles transitions of the topics from one epoch to another.

Bruggermann et al. applied DTM on the RCV1 Reuters corpus (810.000 documents) with weekly time epochs [10]. Results showed that the variances within the topics among the time epochs are marginal. DTM still treats the corpus as a whole and the number of

topics is fixed over all time epochs. Chaney et al. introduce another extension to LDA that detects events in a large text collection [12]. Their model adds separate probability distributions for defined entities and time intervals to the generative process. The model consists accordingly of general topics, entity-related topics and topics specific to time intervals. Events are detected as anomalies, which are identified as temporary deviations from usual behavior. Usual behavior refers to the topics discussed by the entities. An anomaly can be detected, whenever these entities change their topics of discussion significantly and at the same time in a similar way.

3.2.3 Output of LDA. LDA can be applied on a corpus and the output model provides the following distributions:

- $\mathbf{t}_i = \{t_{ij}\}$: topic-word vector-distribution, where i denotes the topic (in total there are N topics) and j denotes the word (in total there are P words in the collection). The component t_{ij} shows the relative weight of word j in topic i.
- $\mathbf{d}_k = \{d_{ki}\}$: document-topic vector-distribution, where i denotes the topic (out of the total N topics) and k denotes a document. The component d_{ki} shows the relative weight of topic i in document k.

4 RESULTS AND ANALYSIS

4.1 Topics in Legal Research on AI (What?)

Pre-processing of the corpus involved filtering for articles written only in the English language and afterwards removing common English stop-words (the, of, etc.) as well as removing some very common corpus-specific words (subject, supra, part, etc.). Moreover, we considered the presence of either unigrams (one words) or bigrams (two words) so as to be able to capture some concepts like 'dispute resolution'.

Subsequently, LDA was used to identify topics in the corpus. For identifying the number of topics (N), we used perplexity [51] and coherence [31] measures on a held-out set. We started from $N = 5$ and increased it (step 5) till $N = 100$. A plateau in both perplexity (178.4) and coherence (0.51) could be observed around 35 or 40 topics, which suggests a computationally optimal number of topics. Based on this, topic models with 30, 35, 40, and 45 topics were explored. Two legal researchers inspected the results of the various topic models in order to determine which number of topics were substantively the most meaningful. For this, the 20 terms with the highest weights for each topic were provided to the researchers (e.g. 'weapon', 'system', 'military', 'international', 'war' etc.). Based on this evaluation, a topic model was selected that consisted of 35 topics.

The topic validation consisted of two steps. First, three researchers inspected the 20 terms for each topic in order to label the topic (e.g. 'military technology'). Second, paper titles were inspected to determine whether the paper titles supported the assigned label - if not, the label would, if possible, be adjusted. In this respect, the researchers were presented with a list of paper titles for each topic. The validation process indicated that the vast majority of the LDA-produced topics were substantively meaningful.

Table 1 reveals the topics distinguished by the LDA model. It includes the topic IDs (not meaningful), the ten terms that have the highest weights in relation to the topic, and the labels the human coders assigned to the topics.

Three miscellaneous topics were identified: id21, id25, and id33. Both the inspection of the 20 words and the titles did not result in a substantively meaningful label or description for these topics. It was decided to not remove the words in these topics and to not re-run the topic modeling algorithm, as it was expected that the removal of words could introduce selection bias in the corpus. Moreover, the identification of the three miscellaneous topics does not affect the relevance or interpretation of the other topics.

The results show a wide range of topics, varying from *tax* to *military technology* to *copyright*. Within each topic, diversity could still be observed. For example, for the topic of *algorithmic decision-making and quantitative methods*, paper titles such as 'An FDA for Algorithms' [49], 'A Simple Guide to Machine Learning' [54], and 'Lawyer as a Soothsayer: Exploring the Important Role of Outcome Prediction in the Practice of Law' [38] can be observed.

An important issue concerned the initial selection of journal articles. The articles were selected based on the search string 'artificial intelligence'. The number of occurrences of this string presumably is, however, not an entirely accurate proxy for measuring whether the paper actually is about artificial intelligence. It might be that papers on *public policy* or *competition & markets* mention the term 'artificial intelligence', but do not primarily focus on artificial intelligence or even on technology or digital matters. An additional selection was therefore required. Consequently, three of the researchers (authors) independently went through the topic list and the related words as displayed in Table 1 in order to determine whether the labels and words are likely to be related to artificial intelligence, digital, and/or technology. A perfect agreement could be observed for the vast majority of topics. In the few instances of disagreement between the coders, the disagreements were resolved through a brief discussion. Ultimately, a total number of 18 topics was selected (Table 1, in bold). The 18 topics were used in subsequent analyses.

4.2 How Did Legal Research on AI Evolve (What - When)?

To answer this question we needed to determine which of the 18 topics that were selected in the previous section, have gained or lost attention over time. The first step is to extract the dominant topics of each paper by using the document-topic vectors d (as described in section 3.2.3). More specifically, we denoted for each document k the first three dominant topics i_{1-3}, based on their relative weights (contributions) in descending order in the document-topic vector $\mathbf{d}_k$. To assess the relevance of each document's topics, two researchers manually reviewed the first five topics – sorted by means of contribution – from a sample of documents. They agreed that for most documents, the relevance dropped significantly after the third topic, since the latter topics provide no substantive contribution to the paper. Based on the results of the inspection, the first three topics were denoted as dominant. We define the *frequency of a topic* as the number of times that a topic is dominant in all articles.

Linking the topic frequencies to the year of publication, the count of papers addressing each topic every year was computed. Figure 1 depicts how the 18 selected topics evolved over time (1960 - 2018).

Table 1: All 35 topics identified by LDA

id	words	labels
0	speech, public, amendment, court, medium, political, free, government, content, freedom	freedom of expression
1	**internet, network, computer, communication, cyberspace, technology, user, access, virtual, service**	**Internet governance**
2	**weapon, system, military, international, war, human, state, attack, target, autonomous_weapon**	**military technology**
3	**datum, information, privacy, personal, protection, data, individual, consumer, big, user**	**privacy**
4	work, company, employee, corporate, worker, labor, business, employer, corporation, economic	labour/corporate
5	lawyer, legal, client, firm, service, practice, attorney, profession, professional, work	legal practice
6	student, school, legal, education, learn, university, skill, practice, teach, research	legal education
7	state, act, agency, public, government, federal, policy, congress, rule, national	public policy
8	**environmental, space, energy, nanotechnology, water, land, risk, plan, air, include**	**environment/space**
9	**patent, claim, invention, method, application, process, art, inventor, court, technology**	**patents**
10	**copyright, work, protection, author, program, copy, court, intellectual_property, fair, computer**	**copyright**
11	**cognitive, behavior, process, theory, make, people, action, social, mental, mind**	**psychology & neuroscience**
12	**human, robot, machine, technology, artificial_intelligence, robotic, agent, science, future, research**	**from humans to machines**
13	market, cost, consumer, service, economic, product, competition, price, trademark, firm	competition & markets
14	**datum, algorithm, model, decision, analysis, result, method, prediction, study, risk**	**algorithmic decision-making & quantitative methods**
15	**surveillance, privacy, government, search, police, enforcement, fourth_amendment, court, information, intelligence**	**surveillance**
16	**contract, party, electronic, agreement, agent, online, term, dispute, transaction, dispute_resolution**	**contract & dispute resolution**
17	legal, theory, social, science, society, system, political, power, economic, form	legal theory/philosophy
18	evidence, probability, argument, theory, inference, case, reason, fact, expert, scientific	evidence
19	international, state, country, european, national, article, trade, member, china, global	international law & relations
20	medical, health, patient, physician, care, medicine, health_care, hospital, fda, device	health
21	https, http, technology, www, online, user, pdf, digital, last_visit, platform	miscellaneous
22	person, human, child, life, moral, legal, animal, state, property, interest	personhood
23	tax, income, trust, taxpayer, property, asset, return, business, pay, interest	tax
24	**legal, rule, case, system, reason, knowledge, base, model, argument, fact**	**knowledge-based systems**
25	time, world, people, game, make, year, life, work, american, story	miscellaneous
26	**financial, market, bank, security, investor, regulation, risk, transaction, investment, trading**	**financial regulation & technology**
27	criminal, crime, police, sentence, justice, offender, sentencing, victim, commit, drug	crime
28	**technology, system, process, change, development, public, information, research, social, design**	**regulation of innovation**
29	**information, search, document, library, legal, research, database, case, access, electronic**	**information retrieval**
30	**liability, vehicle, product, car, tort, autonomous, risk, safety, driver, manufacturer**	**autonomous vehicles**
31	court, case, judge, judicial, rule, justice, decision, opinion, trial, litigation	courts
32	**software, code, license, open_source, source, program, standard, free, developer, computer**	**software licensing**
33	make, rule, problem, case, fact, question, reason, give, decision, view	miscellaneous
34	**computer, system, program, information, software, user, technology, datum, expert, process**	**trends in legal technology**

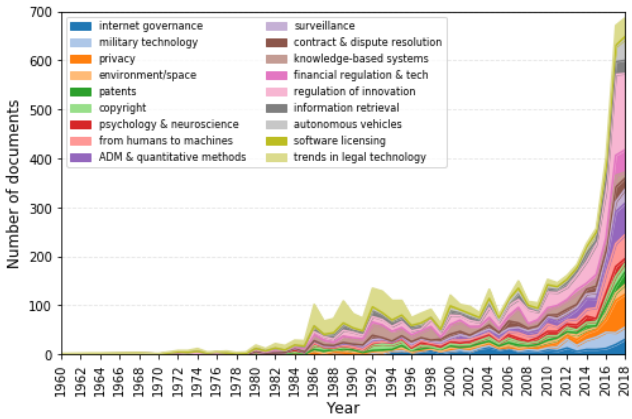

Figure 1: Topic river[1]

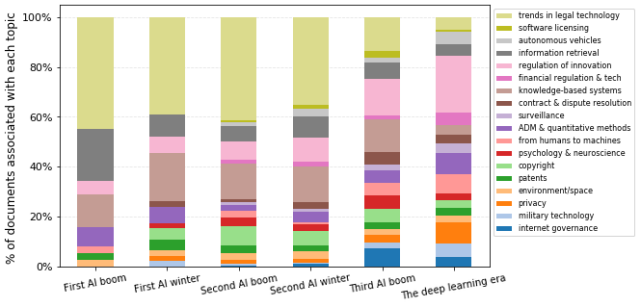

Figure 2: Relative topic evolution across AI periods[2]

Figure 2 shows how the topics have evolved in relation to other topics across six major periods in AI history (taken from [19]).

As a measure of relative topic popularity during a certain period, the ratio of papers concerned with each topic to the total number of papers written in a certain period is computed. Caution needs to be exercised when interpreting the relative topic popularity. Considering the relatively low number of publications before approximately 1986 (see Figure 1), just before the second AI winter, not much weight should be given to the topic popularity in those years, as small changes in the number of publications can have substantial impact on the popularity of one topic relative to other topics.

For example, as visualized in Figure 2, scholarly output produced during the **first AI boom** concerned, among others, *knowledge-based systems* (13.16 %) and *algorithmic decision-making and quantitative methods* (7.90 %), however, *trends in legal technology* and *information retrieval* were prominent topics in this period, forming the subject-matter of 45% and 21% of the papers, respectively. As time advances, topics like *financial regulation & technology*, *autonomous vehicles*, *surveillance*, *software licensing* and *Internet governance* appear.

Another illustration is reflected by the developments visible around the **deep learning era**. The topics that gained popularity (compared to the previous period, namely the **third AI boom**) with rise of deep learning are *regulation of innovation*, *privacy*, *algorithmic decision-making and quantitative methods*, *financial regulation & technology*, *military technology* and *autonomous vehicles*. The largest decreases in popularity in this period are those of *knowledge-based systems* and *trends in legal technology*. The popularity of the other topics underwent popularity changes below the average for this period. The interest in most topics grew during the **deep learning era**.

In addition, we observed several interesting trends regarding the evolution of pairs of topics (e.g. *algorithmic decision-making and quantitative methods* with *knowledge-based systems* and *information retrieval* with *trends in legal technology*), as well as individual topics, e.g. *privacy*, which emerged as a topic during the **first AI winter**, but interest in which saw several relatively insignificant changes before the **deep learning era**, when its popularity underwent a high increase.

[1]The label 'ADM and quantitative methods' refers to the topic *algorithmic decision-making and quantitative methods* (id14).

4.3 Document Similarity

As mentioned in Section 3.2.3, each document k is represented by a vector $\mathbf{d}_k$, where each component of the vector shows the weight of each topic in that specific document. For any pair of documents we can use cosine similarity [42] as a measure of how close two documents are, which in our case is translated to similarity in the 'topic space', i.e. two very similar documents (cosine similarity of value 1) are expected to have similar topic distributions.

The detection of publications that have similar topics enhances information retrieval and reduces the risk of reproduction of scholarship because prior publications are overlooked. Having computed the similarity between all document pairs in our corpus, we obtained some 7,436,296 similarity scores after adjusting the algorithm to avoid computing the similarity between the same document pair twice.

To explore the substantive meaning of the ensuing similarity scores, the results produced by the cosine similarity algorithm were compared to substantive similarity as defined by a legal expert - one of the authors. The goal of the inspection was to explore (1) the extent to which papers were similar and (2) whether differences in similarity scores produced by the cosine similarity measure are substantively meaningful differences. For this, five papers for different topics were selected - the seed papers. Each of these five seed papers were compared with five papers with different similarity scores - the comparison papers: one similarity score in the .55-.65 range, one in the .65-.75 range, one in the .75-.85 range, one in the .85-.95 range, and one in the .95-1.00 range. As a result, a total of five seed papers and 25 comparison papers were inspected.

Without knowledge as to which similarity range the comparison paper would fall under, the legal expert ranked the similarity for each pair of papers (seed paper - comparison paper) for each topic separately. A Spearman's rank correlation test showed a high correlation ($r = .62$, $p < .001$) between the ranks based on the machine-generated similarity scores and the ranks provided by the expert. The comparison took place on four levels: the paper title (i.e. to what extent do the paper titles suggest similarity?), the research question level (i.e. are the research questions similar?), the focus or sub-topic level (i.e. which sub-topics does the paper focus on,

which angle or perspective does it take?), and the citation level (i.e. is there a reference from one paper to the other?).

The inspection of the paper suggests that papers with similarity scores of .85 or lower may share substantive similarity, but in a limited way at best (e.g. pairs of publications where one article [14] focuses on whether (source) code is or should be copyright protected and the other paper [9] on whether results produced by machines are or should be subject to copyright protection). In contrast, articles with high similarity scores, particularly those with a score of .90 or higher, are also substantively similar, at least in case of the inspected papers. For instance, two selected papers on humanitarian law with the highest similarity score ([44] and [45]) discuss legal principles such as proportionality and necessity, and they discuss the role of subjectivity and the capability of autonomous weapons systems (similarity in the >.95 range)

When inspecting citations between papers with very high similarity scores, references from one paper to another were sometimes found, and sometimes not. In some instances when no citation was found this could be due to the fact that papers were published in subsequent years and authors did not have the opportunity to cite the published version of the similar paper. Nevertheless, there are instances where a reference was expected in the papers, but not found.

5 DISCUSSION

The landscape of legal research on AI has undergone considerable changes with respect to the volume of scholarship tackling topics dealing with AI. In this context, we were interested in what exact topics authors focused on, and how these topics shifted over time. For this, we applied LDA topic modeling to identify topics and analyze how journal articles were distributed across topics as well as across time (1960-2018).

The main finding for the first research question, namely what topics can be distinguished in the corpus of legal papers on AI, is that 35 topics can be identified, 32 of them being meaningful (and three miscellaneous topics). Overall, the model performed considerably well in identifying latent topics in our corpus (see Table 1 above).

The second research question dealt with the evolution of topics throughout the different periods which can be identified in the history of AI. The topic river displayed in Figure 1 above shows fundamental changes in legal research on AI from two perspectives: on the one hand, the total number of papers referring to artificial intelligence sees a sharp increase since 2016, and on the other hand, the diversity of topics also increases throughout time. This can be mostly contextualized by the occurrence of new technologies (e.g. the Internet, leading to a new topic on *Internet governance*), but also by the granular development of existing technologies.

As for the third research question, namely how similar papers can be detected, we further calculated similarity scores between pairs of articles and compared the scores for a selection of pairs to the scores produced by humans. Consequently, it was explored whether the similarity scores produced by the machine coincide with the expert assessment. A correlation test revealed a high correlation between the orders produced by the machines and by the human. The highest agreement levels were found for the papers with the highest similarity scores. The similarity predicted by the machine did, however, also sometimes deviate from the expert assessment, although there will undoubtedly also be disagreement between human experts when assessing the similarity of documents.

The results do not only provide insight into how a legal research within a broader theme has evolved over time, they also provide support for the development of LDA tools that assist in structuring large document collections and in finding relevant papers. To aid researchers in either exploring or keeping up with such vast and complex research themes explored in parts of legal scholarship which do not necessarily overlap or interact, it is necessary to consider how a method such as that presented in this paper (topic modeling) can be used to further visualize this body of legal research. To this extent, we are currently working on a dashboard which we aim to make available to scholars (from any discipline) with an interest in exploring the evolution of legal literature on AI, or particular topics within it. Such a dashboard would make it easier, on the one hand, for legal researchers to get the bigger picture of all the fields of law / journals / scholars tackling topics of interest relating to AI, and on the other hand for researchers from other disciplines (e.g. computer science) to have a bird's eye view on the vast legal literature which might have a direct impact on their research. Such publicly available resources could even be a new way to stimulate more awareness of legal and ethical implications of technology on society, and be of interest to civil society as well.

Of course, this paper does not go without limitations. First, the corpus is not necessarily representative of all legal articles or legal publications in general. Although HeinOnline papers do presumably constitute a significant part of journal articles in law, there are other publishers and repositories that contain a number of publications that may composed substantively different than the publications in HeinOnline. Second, another limitation regarding the corpus concerns the initial selection. We selected journal articles that included the keywords 'artificial intelligence'. Had we selected different keywords, the corpus might have looked differently. Additionally, different topics can be expected when running additional topic models for a subset of the corpus. Furthermore, the analyses that explored the increase or decrease of publications over time used the three most dominant topics to determine whether topics have become less or more relevant. The results might be different if also less relevant topics are taken into consideration, although such analyses would be empirically difficult to conduct, as it would require a measure to determine when certain topic weights are deemed insufficiently relevant.

We are currently exploring the possibility of applying and comparing different topic modeling algorithms and incorporate the latest NLP representation models (namely word embeddings either using word2vec or BERT). Moreover, we want to apply our methodology to a different legal collection and identify whether our findings can be confirmed in a different corpus. Finally, after fully validating our approach, we are planning to release it as an open source website where people can explore and visualize our findings in an intuitive way.

REFERENCES

[1] Olufunmilayo Arewa. 2006. Open Access in a Closed Universe: Lexis, Westlaw, Law Schools, and the Legal Information Market. (03 2006).

[2] Claus Boye Asmussen and Charles Møller. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data* 6, 1 (2019), 93. https://doi.org/10.1186/s40537-019-0255-7

[3] Steven Barkan. 1992. Can Law Publishers Change the Law? *Legal Reference Services Quarterly* 11 (03 1992), 29–35. https://doi.org/10.1300/J113v11n03_05

[4] Trevor Bench-Capon, Michal Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald P. Loui, L. Thorne McCarty, Henry Prakken, Frank Schilder, Erich Schweighofer, Paul Thompson, Alex Tyrrell, Bart Verheij, Douglas N. Walton, and Adam Z. Wyner. 2012. A history of AI and Law in 50 papers: 25?years of the international conference on AI and Law. *Artificial Intelligence and Law* 20, 3 (2012), 215–319. https://doi.org/10.1007/s10506-012-9131-x

[5] Robert C. Berring. 1986. Full-Text Databases and Legal Research: Backing into the Future.

[6] Jon Bing. 1986. The text retrieval system as a conversion partner. *International Review of Law, Computers & Technology* 2, 1 (1986), 25–39. https://doi.org/10.1080/13600869.1986.9966228 arXiv:https://doi.org/10.1080/13600869.1986.9966228

[7] David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. ACM, New York, NY, USA, 113–120. https://doi.org/10.1145/1143844.1143859

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[9] A. Bridy. 2012. Coding Creativity: Copyright and the Artificially Intelligent Author. *Stanford Technology Law Review* 5 (2012), 1–28.

[10] Daniel Brüggermann, Yannik Hermey, Carsten Orth, Darius Schneider, Stefan Selzer, and Gerasimos Spanakis. 2016. Storyline detection and tracking using Dynamic Latent Dirichlet Allocation. *Computing News Storylines* (2016), 9.

[11] David J Carter, James Brown, and Adel Rahmani. 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. *UNSWLJ* 39 (2016), 1300.

[12] Allison June-Barlow Chaney, Hanna M Wallach, Matthew Connelly, and David M Blei. 2016. Detecting and Characterizing Events.. In *EMNLP*. 1142–1152.

[13] Laura Dietz, Bhaskar Mitra, Jeremy Pickens, Hana Anber, Sandeep Avula, Asia J. Biega, Adrian Boteanu, Shubham Chatterjee, Jeff Dalton, Shiri Dori-Hacohen, John Foley, Henry Feild, Ben Gamari, Rosie Jones, Pallika Kanani, Sumanta Kashyapi, Widad Machmouchi, Matthew Mitsui, Steve Nole, Alexandre Tachard Passos, Jordan Ramsdell, Adam Roegiest, David Smith, and Alessandro Sordoni. 2019. Report on the First HIPstIR Workshop on the Future of Information Retrieval. *SIGIR Forum* 53, 2 (December 2019), 62–75. https://www.microsoft.com/en-us/research/publication/report-on-the-first-hipstir-workshop-on-the-future-of-information-retrieval/

[14] R. Dixon. 2003. Breaking into locked rooms to access computer source code: Does the dmca violate constitutional mandate when technological barriers of access are applied to software. *Virginia Journal of Law & Technology* 8, 1 (2003), 1–60.

[15] Arthur Dyevre and Nicolas Lampach. 2018. Issue Attention on the European Court of Justice: A Text-Mining Approach. *SSRN* (2018). http://dx.doi.org/10.2139/ssrn.3251186

[16] Ian Edwards. 2018. Search like a robot: Developing targeted search algorithms. *Australian Law Librarian* 26, 2 (2018), 104.

[17] Daphne Gelbart and J. Smith. 1993. Automating the Process of Abstracting Legal Cases. *International Journal of Law and Information Technology* 1 (03 1993), 324–334. https://doi.org/10.1093/ijlit/1.3.324

[18] Daphne Gelbart and J. C. Smith. 1994. The application of automated text processing techniques to legal text management. *International Review of Law, Computers & Technology* 8, 1 (1994), 203–210. https://doi.org/10.1080/13600869.1994.9966390 arXiv:https://doi.org/10.1080/13600869.1994.9966390

[19] Catalina Goanta, Gijs van Dijck, and Gerasimos Spanakis. 2019. Back to the Future: Waves of Legal Scholarship on Artificial Intelligence. *Forthcoming in Sofia Ranchordás and Yaniv Roznai, Time, Law and Change (Oxford, Hart Publishing, 2019)* (2019).

[20] Graham Greenleaf, Daniel Austin, Philip Chung, Andrew Mowbray, Madeleine Davis, and Jill Matthews. 2000. Solving the Problems of Finding Law on the Web: World Law and DIAL. *Journal of Information, Law and Technology* 2000 (01 2000). https://doi.org/10.1017/S0731126500009483

[21] Graham Greenleaf, Andrew Mowbray, Geoffrey King, and Geoffrey van Dijk. 1995. Public Access to Law via Internet: The Australian Legal Information Institute. *Journal of Law and Information Science, 6(1), 50* (1995).

[22] F. Hanson. 2002. From Key Numbers to Keywords: How Automation Has Transformed the Law. *Law Library Journal* 94 (09 2002).

[23] Karen Hao. 2019. We analyzed 16,625 papers to figure out where AI is headed next. https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/

[24] Maria Iglesias, Sharon Shamuilia, and Amanda Anderberg. 2019. Intellectual Property and Artificial Intelligence - A literature review. *EU Science Hub - European Commission* (Dec. 2019). https://ec.europa.eu/jrc/en/publication/intellectual-property-and-artificial-intelligence-literature-review

[25] Aaron Kirschenfeld. 2017. Yellow Flag Fever: Describing Negative Legal Precedent in Citators. https://doi.org/10.31228/osf.io/dfjah

[26] Melanie Knapp and Rob Willey. 2016. Comparison of Research Speed and Accuracy Using WestlawNext and Lexis Advance. *Legal Reference Services Quarterly* 35 (05 2016), 1–11. https://doi.org/10.1080/0270319X.2016.1177428

[27] Michael A Livermore, Allen Riddell, and Daniel Rockmore. 2016. Agenda formation and the US supreme court: A topic model approach. *Arizona Law Review* 1, 2 (2016).

[28] Peter Maggs. 1994. Legal Data Banks in the United States and Their Use in Comparative Law. *International Journal of Legal Information* 22 (01 1994), 214–227. https://doi.org/10.1017/S0731126500024926

[29] Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. 2018. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures* 12, 2-3 (2018), 93–118. https://doi.org/10.1080/19312458.2018.1430754 arXiv:https://doi.org/10.1080/19312458.2018.1430754

[30] Elizabeth M. McKenzie. 2001. Natural Language Searching. *Legal Reference Services Quarterly* 18, 4 (2001), 39–47. https://doi.org/10.1300/J113v18n04_04 arXiv:https://doi.org/10.1300/J113v18n04$_0$4

[31] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.

[32] A Moens. 2009. Is AustLII Sustainable? *Australian Law Librarian,17(3), 154-157* (2009).

[33] Marie-Francine Moens, Maarten Logghe, and Jos Dumortier. 2002. Legislative Databases: Current Problems and Possible Solutions. *International Journal of Law and Information Technology* 10 (03 2002). https://doi.org/10.1093/ijlit/10.1.1

[34] Elizabeth Moll-Willard. 2018. The use and perceptions of open Access resources by legal academics at the University of Cape Town (UCT) in South Africa. 6 (09 2018), 1–13.

[35] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 215–224.

[36] P. Ogden. 1993. "Mastering the lawless science of our law": A story of legal citation indexes. *Law Library Journal* 85 (01 1993), 1–48.

[37] Marc Opijnen, Hayo Schreijer, Ilja Andreas, and Maarten Kroon. 2015. Specialised Government Publishing: The Law Pocket and Linked Legal Data in the Netherlands.

[38] Mark K. Osbeck. 2018. Lawyer as Soothsayer: Exploring the Important Role of Outcome Prediction in the Practice of Law. *Penn State Law Review* 123, 1 (2018), 41–102.

[39] Yannis Panagis and Urska Sadl. 2015. The Force of EU Case Law: A Multidimensional Study of Case Citations.. In *JURIX*. 71–80.

[40] João Reis, Paula Espírito Santo, and Nuno Melão. 2019. Impacts of Artificial Intelligence on Public Administration: A Systematic Literature Review. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 1–7.

[41] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[42] G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM* 18, 11 (Nov. 1975), 613–620. https://doi.org/10.1145/361219.361220

[43] Pamela Samuelson. 2010. Google Book Search and the Future of Books in Cyberspace. *Minnesota law review* 94 (01 2010).

[44] M. Sassoli. 2014. Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified. *International Law Studies. US Naval War College* 90 (2014), 308–340.

[45] Michael Schmitt and Jeffrey Thurnher. 2013. 'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict. *Harvard National Security Journal* 4, 02 (2013), 231–281.

[46] Cecilia Magnusson Sjóberg. 1997. Corpus Legis: A Legal Document Management Project. *International Journal of Law and Information Technology* 5, 1 (03 1997), 83–99. https://doi.org/10.1093/ijlit/5.1.83 arXiv:https://academic.oup.com/ijlit/article-pdf/5/1/83/9820642/83.pdf

[47] James A. Sprowl. 1976. Computer-Assisted Legal Research—An Analysis of Full-Text Document Retrieval Systems, Particularly the LEXIS System. *Law & Social Inquiry* 1, 1 (1976), 175–226. https://doi.org/10.1111/j.1747-4469.1976.tb00955.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1747-4469.1976.tb00955.x

[48] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational

Linguistics, 952–961.
[49] Andrew Tutt. 2017. An FDA for Algorithms. *Administrative Law Review* 69, 1 (2017), 83–123.
[50] Marc van Opijnen. 2017. Gaining Momentum. How ECLI Improves Access to Case Law in Europe.
[51] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*. 1105–1112.
[52] Chong Wang, David M. Blei, and David Heckerman. 2008. Continuous Time Dynamic Topic Models.. In *UAI*, David A. McAllester and Petri Myllymäki (Eds.). AUAI Press, 579–586. http://dblp.uni-trier.de/db/conf/uai/uai2008.html#WangBH08
[53] Xuerui Wang and Andrew McCallum. 2006. Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA) *(KDD '06)*. ACM, New York, NY, USA, 424–433. https://doi.org/10.1145/1150402.1150450
[54] E. Warren. 2017-2018. A Simple Guide to Machine Learning. *SciTech Lawyer* 14, 1 (2017-2018), 5–9.
[55] Clemens Wass. 2017. openlaws.eu – Building Your Personal Legal Network.
[56] Xing Wei, Jimeng Sun, and Xuerui Wang. 2007. Dynamic Mixture Models for Multiple Time-Series.. In *IJCAI* (2007-03-05), Manuela M. Veloso (Ed.). 2909–2914. http://dblp.uni-trier.de/db/conf/ijcai/ijcai2007.html#WeiSW07
[57] Robin Widdison. 2002. New Perspectives in Legal Information Retrieval. *International Journal of Law and Information Technology* 10, 1 (01 2002), 41–70. https://doi.org/10.1093/ijlit/10.1.41 arXiv:https://academic.oup.com/ijlit/article-pdf/10/1/41/2065390/100041.pdf
[58] R Winkels. 2015. Experiments in finding relevant case law. In *NAIL 2015: 3rd International Workshop on Network Analysis in Law.*

A APPENDIX

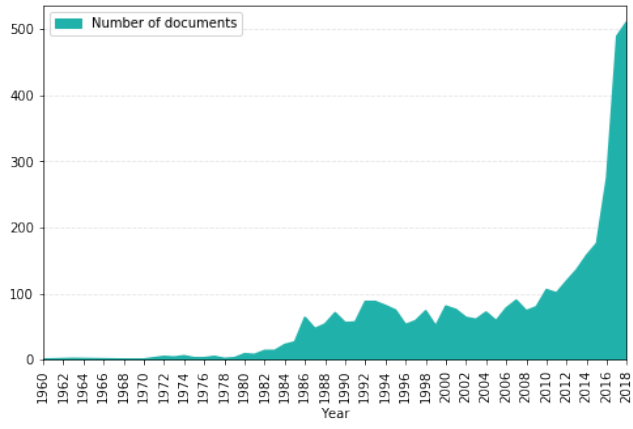

Figure 3: Number of journal articles with keywords per year

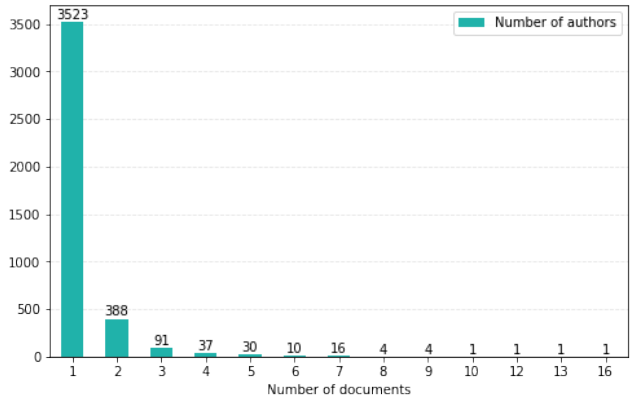

Figure 4: Distribution of the number of authors over the number of documents

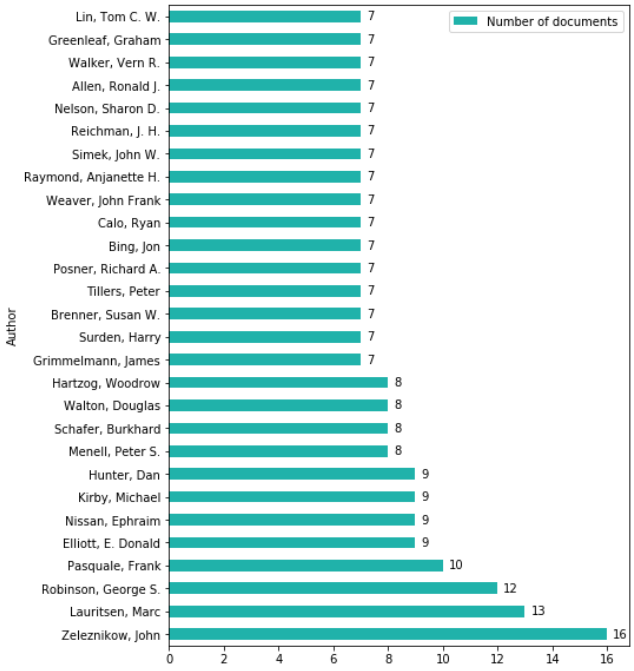

Figure 5: Number of publications per author for authors with at least five publications