Topic Modelling on the European AI Act

Ettore Carbone[1,*], Alex Giulio Berton[1], Purbasha Chowdhury[1], Teresa Scantamburlo[2,3] and Paolo Falcarin[1]

[1]Ca' Foscari University of Venice, Italy
[2]University of Trieste, via Economo 12/3, 34123 Trieste
[3]European Centre for Living Technology, Ca' Bottacin, Dorsoduro 3911, Calle Crosera, 30123 Venice, Italy

Abstract

This work explores knowledge acquisition and representation tools for automatically creating a high-level model representation of the European regulation on artificial intelligence, commonly known as AI ACT. We utilized BERTopic for extracting topics and we also focused on the comparative analysis with other language models based on the topic extractions and representations. Natural language processing and comprehension of legal text is becoming important as legal texts are often interconnected with a large number of other related materials. Therefore, legal text analysis requires technologies which are able to extract important topics and representing them into a comprehensive form, in order to correctly inform the requirements engineering process.

Keywords

Requirements Elicitation, Legal Compliance, Information extraction, Topic Modelling, AI ACT

1. Introduction

The EU digital strategies are becoming very important for software engineers as they concern any software services offered in the European Union. Nowadays, people and businesses outside the EU may be subject to General Data Protection Regulation (GDPR) obligations and this is also the case with the Digital Markets Act (DMA), Digital Services Act (DSA), and the Artificial Intelligence Act (AI Act).

The sanctions for non-compliance are significant: they are frequently expressed as a percentage of annual global revenues (up to 10% in the case DMA, 6% under the DSA and draft AI Act, and 2.5% under the proposed Cyber Resilience Act). Therefore, software systems that interact with personal or sensitive information must be designed with compliance in mind from the outset.

In order to simplify the understanding of legal requirements it is useful to summarize them into a higher-level representation that can highlight the important concepts and relationships into a sort of reusable knowledge.

In this work, we are applying topic modelling on the The Artificial Intelligence Act [1], also known as the AI Act i.e., the world's first statutory law proposal for regulating AI systems. The AI Act aims to ensure that AI systems in the EU are safe and respectful of fundamental rights and values. Its application extends beyond EU-based organizations and regards any AI provider, importer, distributor, or authorized representative within the EU. The expected impact of the AI Act is enormous [2], and the United Nations moves towards a globally coordinated AI governance [3], as AI raises concerns not only for data protection but also transparency [4], fairness [5], information privacy, and freedom of expression [6].

The paper is structured as follows: Section II sets the background, presenting the AI Act's risk based approaches and the related works. Section III presents the proposed topic modelling methodology, Section IV describe the results, while Section V concludes the paper with some future directions.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.
ettore.carbone@unive.it (E. Carbone); 884378@stud.unive.it (A. G. Berton); purbasha.chowdhury@unive.it (P. Chowdhury); teresa.scantamburlo@units.it (T. Scantamburlo); paolo.falcarin@unive.it (P. Falcarin)
0009-0004-7386-6910 (E. Carbone); 0009-0003-4022-9766 (P. Chowdhury); 0000-0002-3769-8874 (T. Scantamburlo); 0000-0003-1933-5348 (P. Falcarin)

2. Background

Our aim is using topic modelling on a legal text, with particular focus on EU directives that are affecting software systems and services, and we used the EU AI Act as a case study, as recent works on AI Act compliance showed the limits of manual elicitation and validation of requirements from a huge corpora of legal text [7].

2.1. The AI ACT

The AI Act is part of a broader EU strategy designed to enhance Europe's potential to compete globally in regulating the digital sector. A number of proposed laws would establish regulatory bodies at the EU and member state level, with broad investigative and enforcement powers, including an European Artificial Intelligence Board to oversee the AI Act, while a European Board for Digital Services is responsible for the Digital Services Act.

The AI Act proposal delineates four distinct risk categories and sets specific requirements accordingly. These categories are:

- Unacceptable Risk (Title II - Art. 5 and following);
- High Risk (Title III - Art. 6 and following);
- Limited Risk (Art. 52);
- Minimal Risk / No Risk;

Companies developing or deploying high-risk AI systems must comply with various requirements, including having an appropriate risk management system, logging capabilities, and human oversight (see Chapter 2 of the AI Act). For systems deemed to pose an unacceptable risk, which is outright prohibited, the Act provides explicit examples and exceptions, including the utilization of real-time remote biometric identification in public spaces (such as facial recognition), social scoring systems (classifying individuals based on behaviour, socio-economic status, or personal characteristics), and the use of subliminal manipulation techniques targeting specific vulnerable groups (Art. 5).

High-risk systems are permitted, but due to their ability to negatively affect safety or fundamental rights, they must comply with multiple requirements and undergo a compliance assessment throughout their life cycle, including before and after being deployed. High-risk systems are further divided into two categories (Art. 6, Annexes II and III):

- AI systems intended to serve as safety components in products covered by the legislation listed in Annex II, or subject to third-party ex-ante conformity assessment (e.g., toys, aviation, cars, medical devices, and lifts).
- Stand-alone AI systems with mainly fundamental rights implications, listed in Annex III, that will have to be registered in an EU database.

Examples of high-risk systems include those related to critical infrastructure management, systems in hiring processes or employee ratings, credit scoring systems, and systems with critical impact on law enforcement and interpretation of law. The recent division of AI Act includes the AI systems into three main risk categories for compliance purposes: Unacceptable, High, and General Purpose AI (GPAI). Originally, Limited and Minimal Risk were separate categories, but the final version affiliates these into general guidelines with minimal obligations. The compromise agreement dated 26th January 2024 formally introduces GPAI models (Articles 52a–52e), whose risk-level needs to be assessed on a case-by-case basis.

Similar to the GDPR, proper data governance must be applied to users' (and, more broadly, data subjects') data, but, in comparison to the data protection regulatory framework, the AI Act goes further by requiring data governance for data used in the training, testing, and validation of AI systems.

2.2. Related work

Topic modelling is used to extract common themes or to cluster similar documents, often serving as a foundation for more complex NLP tasks in the legal domain, such as retrieving similar cases or classifying legal documents. It can also assist legal experts in the annotation of legal texts by providing a preliminary grouping of related materials, thereby improving efficiency for a labour and time-intensive task. Alternatively, topic modelling can be used to enhance datasets manually curated by experts, leading to improved accuracy in document classification [8]. It is also a relevant methodology to automatically annotate metadata as demonstrated by Tuarob et al. [9].

More broadly, topic modelling contributes to ongoing efforts to leverage NLP techniques for the analysis and management of legal texts (for a complete overview of the field see this survey: [10]), or what has been called Legal AI [11]. Topic modelling of legal texts can be done in various ways. A standard approach is based on Latent Dirichlet Allocation (LDA) techniques [12, 13], a generative approach that models words in documents as being probabilistically sampled from underlying latent variables [14].

A more recent approach is based on clustering approaches over vector representation of the text (*embeddings*), where the vectors are created using transformer-based models [15, 16]. Cabot et al. [17] have proposed a method to verify GDPR compliance in data processing agreements using NLP to compare contractual language with mandatory legal provisions. Similarly, Lippi et al. [18] introduced a new framework that uses machine learning and NLP techniques to evaluate legal compliance by automatically extracting processes for obligations and constraints from the regulatory texts. Natural Language Processing (NLP) approaches to the analysis of legal documents have largely relied on topic modelling. Motivated by the overabundance of often interlinked materials, this technique has been applied across various pieces of legislation ranging from court decisions [19] to national and regional statutory laws [20, 21].

A recent study investigated the effectiveness of combining topic modelling techniques with contextualized embeddings and various preprocessing strategies to organize and analyse large collections of Brazilian legal documents across diverse formats and lengths [22]. Another related study applies topic modelling to improve the semantic retrieval and summarization of court judgments, enabling more effective matching of user queries and highlighting the most relevant content to enhance search efficiency and user understanding [23].

Our work applies and compares topic modelling techniques to the analysis of the European AI Act, with the aim of exploring how extracted topics could be used to represent the legal text and, possibly, support preliminary phases of legal compliance. For example, topic modelling could serve as a foundation for creating or extending legal ontologies or for building structured representations of legal texts that support machine-readable compliance frameworks, similar to approaches used for cybersecurity regulation in IoT governance [24].

3. Methodology

In our work, the pipeline adopted to retrieve the topic from the whole legal text is divided into the following two main steps:

1. Segmentation: The text was segmented into shorter passages to serve as input for topic modelling.
2. topic modelling: We applied a topic modelling approach to cluster similar chunks of the legal text and present them with a human-understandable representation.

For the segmentation step, we applied three different approaches in order to produce different topic representations of the whole original text.

We design our experiments to mimic what an average skilled practitioner might want to apply to create a topical representation of a legal text, and therefore we employ techniques that are widely adopted among the machine-learning community and that are easily accessible. In particular, to perform

topic modelling we applied: Latent Dirichlet Allocation (LDA) [14], BERTopic [25]; and Large Language Models (LLMs) [26] prompting technique, specifically ChatGPT 4o[27].

3.1. Text segmentation

To prepare the dataset for topic modelling, we split the text of the AI Act into small chunks from the official web page of the legislation [1]. By examining the HTML structure of the document, we devised a hierarchical division, where nested tags represented different levels of paragraphs. First we separated the recitals, enacting terms, and annexes. When present, we further divided the content into chapters and sections based on the document structure. The individual articles were easily identifiable, as they were enclosed in `div` tags, and their internal structure followed three distinct levels of depth using `table` and `p` tags. The final output was a table in which each row contained a paragraph along with its corresponding metadata (e.g. the part and article it belongs to) and a unique identifier. We chose the numbered paragraph as the unit of our topic analysis, as it represents a coherent piece of text conveying self-contained semantics. A summary of the dataset's statistics is presented in Table 1.

Table 1
Relevant statistics of the segmented legal text.
The average, minimum, and maximum length refer to the numbered paragraphs.

Part	Articles	Paragraphs	Avg. Length	Min Length	Max Length
1	180	1810	1254.97	140	4043
2	113	634	447.4	8	4650
3	13	150	286.52	5	2705

3.2. Topic modelling

We assume that the segmentation technique employed outputs a set of paragraphs $\{p_i\}_{i=0}^{P}$, where p_i represents the i-the paragraph, and P is the total number of paragraphs; in our case, $P = 2594$. Note that these correspond to the numbered paragraphs in the text of the AI Act. We assume that each p_i can be associated with at least one topic, and to allow the comparison between different topic modelling strategies, we further assume that each paragraph p_i can be broadly represented with only one topic t_i, with $1 \leq i \leq T$, where T is the total number of topic present in the collection of paragraphs. Thus, we define a topic modelling technique as a function M which, given a paragraph p_i, returns a topic t_i, i.e., $M(p_i) = t_i$. In the following subsections, we provide a brief description of the three methods adopted.

LDA LDA is a standard technique used to perform topic modelling presented for the first time by Blei et al. [28] in 2001, and it is still considered one of the main baselines to be used in a comparative analysis and a vast literature of applications and improved implementation has been developed during recent years [29].

LDA, using a standard statistical terminology, is a mixture model, in which each document (or paragraph) is a mixture of topics, and each topic is a mixture of words. The model is usually described as a Bayesian Network, and in our experiments, we adopted an online variational Bayes algorithm [30] to approximate the target posterior distributions using the implementation provided by scikit-learn [31].

It should be noted that LDA does not associate a single topic with each paragraph, but instead, a probability distribution of topics is associated. However, to allow a better comparison with the other strategies, we pick the topic with the highest probability as the topic associated with the paragraph. Formally, if $LDA(p_i, t_j)$ is the LDA function that estimates the probability that the topic t_j is present in a specific paragraph p_i. Thus, given a paragraph p, topic model based on LDA M_{LDA} associates the

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

topic with the highest probability among all the available topic:

$$M_{LDA}(p) = \arg\max_{t_i} LDA(p, t_j) \quad \text{with} \quad 1 \leq j \leq T$$

BERT Topic The BERT Topic[25] approach is instead based on the assumption that the vector representation produced by an encoder-only pre-trained language model as BERT, can represent the paragraphs semantically in the vector space created at the end of the encoding phase.

In particular, BERT Topic assigns topics to a set of paragraphs employing the following steps:

1. The paragraphs are encoded into the vector space (also called embedding space).
2. The embeddings are down-projected to a lower-dimensional space.
3. The down-projected embeddings are clustered together using a specific similarity measure between vectors.
4. A representation for each cluster is provided to the end user.

For the first step we used the default SBERT model to generate the embeddings of the documents. For the second step we decided to use *CountVectorizer* from *scikit-learn* removing English stop words. For the third step, we used K-means, experimenting with different numbers of clusters and selecting the one that yielded the highest in-group similarity.

LLM Prompting The last strategy we explored leverages the direct use of a large language model (LLM) to assign topics to each paragraph through prompt engineering. In this case, we employed ChatGPT-4o. The motivation for adopting this approach lies in its accessibility: unlike traditional topic modelling algorithms, which often require technical expertise for data preparation, parameter tuning, and model evaluation, an LLM-based solution can, in principle, be used by any layperson without specialized knowledge. Through a simple natural language prompt, non-experts can obtain a high-level overview of the topics addressed in the AI Act, as well as identify where these topics are discussed across the text. So in our research we wanted to simulate this possibility and compare it with the other methods. For this experiment, we provided ChatGPT-4o with the following instruction: "Perform topic modelling of the column 'text' of the attached data file in CSV format and provide the result of the topic modelling in a CSV format including topic number, keywords, and number of documents associated." Notably, we did not supply any example of the expected output format but the model successfully produced the requested results in the desired structure.

4. Experimental evaluation

All experiments were conducted using the Python programming language, employing several libraries to facilitate model implementation and evaluation. For the BERTopic-based approach, we utilized the BERTopic library, with *KMeans* as the clustering algorithm and *CountVectorizer* (both from the *scikit-learn* library) employed for vectorization, removing English stop words from the input.

For the Latent Dirichlet Allocation (LDA) model, we also relied on the implementation provided by *scikit-learn*. Hyperparameters such as the maximum number of iterations and the learning method were optimized through empirical testing across multiple runs.

To determine the optimal number of topics for both BERTopic and LDA, we developed a heuristic algorithm that systematically evaluated cluster sizes ranging from 5 to 10. For each candidate number of clusters, we computed a cohesion score based on the average pairwise similarity of the topic keywords. Specifically, we encoded the keywords using *Sentence Transformer* embeddings and calculated their cosine similarity, subsequently averaging these scores for each topic and across all topics at a given cluster size. The number of topics yielding the highest average similarity score was selected, as it indicates more internally coherence.

In the case of ChatGPT-4o, no constraints were imposed on the number of topics. The large language model was prompted directly and allowed to determine both the number of topics and their composition

based on its internal understanding of the text. Following topic assignment at the paragraph level, post-processing was performed using the *Pandas* library to aggregate the results by topic, enabling consistent comparison with the outputs generated by the other methodologies.

4.1. Analysis of results

We decided to present results as top ten keywords per topic as shown in Table II.

Table 2
Top 10 keywords for each topic by the different methodology implemented

	BERT	**LDA**	**GPT**
Topic 1	ai, systems, highrisk, union, regulation, persons, shall, use, data, law	market, model, provider, general, high, systems, shall, purpose, risk, ai	model, data, purpose, general, models, training, including, used, information, content
Topic 2	shall, notified, body, assessment, conformity, commission, bodies, member, provider, requirements	biometric, behalf, natural, content, ai, offences, enforcement, law, criminal, person	systems, risk, high, regulation, testing, providers, including, market, union, relevant
Topic 3	surveillance, authorities, shall, market, regulation, testing, ai, authority, regulatory, conditions	techniques, objectives, inputs, outputs, rights, data, content, systems, text, ai	regulation, shall, union, european, authorities, commission, article, national, systems, member
Topic 4	ai, model, generalpurpose, models, office, systemic, including, information, ageneralpurpose, shall	intended, use, data, shall, used, persons, high, risk, systems, ai	systems, persons, data, law, use, natural, used, biometric, identification, personal
Topic 5	eu, council, parliament, european, regulation, oj, delegated, acts, shall, commission	whichever, higher, preceding, turnover, worldwide, trafficking, illicit, tfeu, eur, 000	shall, article, risk, referred, provider, high, notified, market, conformity, assessment
Topic 6	biometric, identification, use, data, law, remote, systems, enforcement, realtime, purpose	carrying, activities, type, entity, regulation, military, national, defence, security, purposes	-
Topic 7	-	enforcement, shall, administrative, authorisation, systems, remote, law, biometric, identification, use	-
Topic 8	-	assessment, surveillance, body, member, authority, commission, referred, notified, article, shall	-
Topic 9	-	2008, 2014, european, 2017, repealing, parliament, amending, council, ec, oj	-
Topic 10	-	authorities, systems, law, european, union, ai, data, article, regulation, eu	-

The distribution of topics for each methodology is shown in *Figure 1*. The three approaches produced a different number of topics: BERTopic generated 6 topics, LDA produced 10 topics, and ChatGPT identified 5 topics. As can be observed, the BERTopic model produces an almost uniform distribution across topics, with the exception of Topic 1, which contains a significantly higher number of paragraphs compared to the others. ChatGPT yields similar results, although two of its topics include more than 250 paragraphs each. Conversely, the LDA model generates the highest number of topics, resulting in a more uneven distribution: Topic 1 includes approximately 300 paragraphs, while Topics 5 and 6 contain fewer than 10 paragraphs each.

A preliminary qualitative analysis of the identified topics and their corresponding clustered paragraphs revealed mixed results. The topic modelling outcomes were manually examined and discussed in meetings among the authors. A more in-depth analysis (e.g., through coding and thematic analysis) is planned for future work to obtain more fine-grained insights into the results of the topic modelling techniques. Qualitative analysis suggests that, in some cases, the topic keywords were well-aligned with the content of the assigned paragraphs, indicating meaningful and coherent clusters. For example, in the case of Topic 6 from the LDA division, the content of paragraph 3 of article 2 shows high coherence with the identified keywords. However, for other topics, the clustered paragraphs were considerably more heterogeneous, suggesting that the models may have grouped together paragraphs covering

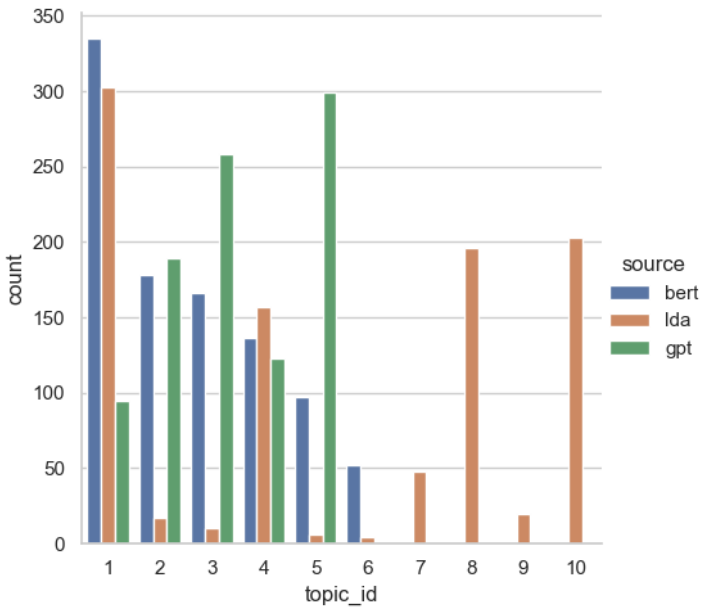

Figure 1: The distribution of paragraphs for each topic found by the various methodologies (note that the number of topics varies among them).

diverse or loosely related content under a single topic label. For instance, recital 98 states: "Large generative AI models are a typical example of a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks." This paragraph was grouped under Topic 5, even though none of the top keywords appeared in the text and its content lacked coherence with the other documents assigned to that topic.

5. Conclusion and Future work

In this paper, we propose a method to perform topic mining on the legal articles of the AI Act European Directive. We used BERT, LDA and GPT methods to do topic analysis, with the goal of analysing legal requirements using a higher level representation. Our goals were also to compare different topic modelling approaches in order to test their behaviour on the same task and assess their limits. Building on approaches such as those proposed by Hagen et al.[13], domain experts could assess both the quality of the generated topic keywords and the coherence between these keywords and the associated paragraphs.

Qualitative analysis has been used to create formal taxonomies out of cybersecurity reports in natural language [32]. Similarly, automated topic modelling could support the development of such formal representations for cybersecurity [33], but also for legal texts, such as the AI Act. This hybrid approach may serve as a premise for constructing a structured knowlede sege base that facilitates systematic exploration of the regulation's thematic content. Additionally, topic modelling aims to be applied on other legal directives such as NIS2, cyber-resilience etc. as well as software requirement specifications of different enterprises. The methodologies of this work could also be employed as groundworks for other tasks, such as semantic search as proposed by Ma et al. [34], summarization as done by Haghighi et al. [35], or as a guiding tool in the construction of knowledge graphs from the AI ACT and other legal directives [36], in order to improve the translation from unstructured text to data model for requirement compliance analysis.

Acknowledgments

The authors want to thank Dr Alberto Veneri for his useful insights and for reviewing our work.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] European Union, Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[2] S. Bertaina, I. Biganzoli, R. Desiante, D. Fontanella, N. Inverardi, I. G. Penco, A. C. Cosentini, Fundamental rights and artificial intelligence impact assessment: A new quantitative methodology in the upcoming era of AI act, Computer Law & Security Review 56 (2025) 106101.

[3] UN Office of the Secretary-General's Envoy on Technology, High-level advisory body on AI, 2023. URL: https://www.un.org/techenvoy/ai-advisory-body.

[4] J. Burrell, How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society 3 (2016) 2053951715622512. doi:10.1177/2053951715622512.

[5] S. Barocas, A. Selbst, Big data's disparate impact, SSRN Electronic Journal (2016). doi:10.2139/ssrn.2477899.

[6] T. Scantamburlo, A. Charlesworth, N. Cristianini, Machine decisions and human consequences, CoRR abs/1811.06747 (2018). URL: http://arxiv.org/abs/1811.06747. arXiv:1811.06747.

[7] T. Scantamburlo, P. Falcarin, A. Veneri, A. Fabris, C. Gallese, V. Billa, F. Rotolo, F. Marcuzzi, Software systems compliance with the AI Act: Lessons learned from an international challenge, in: Proc. of the 2nd International Workshop on Responsible AI Engineering, RAIE '24, ACM, 2024, p. 44–51. doi:10.1145/3643691.3648589.

[8] L. J. G. Freitas, T. Rodrigues, G. Rodrigues, P. Edokawa, A. Farias, Text clustering applied to data augmentation in legal contexts, arXiv preprint arXiv:2404.08683 (2024).

[9] S. Tuarob, L. C. Pouchard, P. Mitra, C. L. Giles, A generalized topic modeling approach for automatic document annotation, International Journal on Digital Libraries 16 (2015) 111–128.

[10] F. ARIAI, G. DEMARTINI, Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges, ACM Comput. Surv 1 (2024).

[11] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Ass. for Computational Linguistics, Online, 2020, pp. 5218–5230. URL: https://aclanthology.org/2020.acl-main.466/.

[12] A. Mandal, R. Chaki, S. Saha, K. Ghosh, A. Pal, S. Ghosh, Measuring similarity among legal court case documents, in: Proceedings of the 10th annual ACM India compute conference, 2017, pp. 1–9.

[13] L. Hagen, Content analysis of e-petitions with topic modeling: How to train and evaluate lda models?, Information Processing & Management 54 (2018). doi:10.1016/j.ipm.2018.05.006.

[14] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[15] R. Silveira, C. G. Fernandes, J. Araujo Monteiro Neto, V. Furtado, J. E. Pimentel Filho, Topic modelling of legal documents via Legal-BERT, in: Proc. of the 1st Intern. Workshop RELATED - Relations in the Legal Domain 2021, 2021. doi:10.2139/ssrn.4539091.

[16] H. Sargeant, A. Izzidien, F. Steffek, Topic classification of case law using a large language model and a new taxonomy for uk law: Ai insights into summary judgment, Artificial Intelligence and Law (2025) 1–49.

[17] P.-L. H. Cabot, R. Navigli, Rebel: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2370–2381.

[18] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, P. Torroni, Claudette: an automated detector of potentially unfair clauses in online terms of service, Artificial Intelligence and Law 27 (2019) 117–139.
[19] B. J. U. Razon, G. A. Solano, L. T. B. Ranera, Topic modelling supreme court case decisions using latent dirichlet allocation, in: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2022, pp. 284–289.
[20] J. O'Neill, C. Robin, L. O'Brien, P. Buitelaar, An analysis of topic modelling for legislative texts, in: Proc of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL'17), CEUR Workshop Proceedings, London, UK, 2017.
[21] A. Dyevre, M. Glavina, M. Ovádek, The voices of european law: legislators, judges and law professors, German Law Journal 22 (2021) 956–982.
[22] D. Vianna, E. S. de Moura, A. S. da Silva, A topic discovery approach for unsupervised organization of legal document collections, Artificial Intelligence and Law 32 (2024) 1045–1074.
[23] T.-H. Wu, B. Kao, F. Chan, A. S. Cheung, M. M. Cheung, G. Yuan, Y. Chen, Semantic search and summarization of judgments using topic modeling, in: Legal Knowledge and Information Systems, IOS Press, 2021, pp. 100–106.
[24] S. S. Chennu, L. Elluri, G. Batra, Bridging AI and legal compliance: Knowledge graphs for IoT cybersecurity regulations, in: AMCIS 2025 Proceedings, 22, Assoc. for Information Systems, 2025.
[25] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
[26] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (2024). doi:10.1145/3641289.
[27] OpenAI, Hello GPT-4o, https://openai.com/index/hello-gpt-4o/, 2025. Accessed: 2025-06-10.
[28] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, Advances in neural information processing systems 14 (2001).
[29] U. Chauhan, A. Shah, Topic modeling using latent dirichlet allocation: A survey, ACM Comput. Surv. 54 (2021). doi:10.1145/3462478.
[30] M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference, J. Mach. Learn. Res. 14 (2013) 1303–1347.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011).
[32] M. Ceccato, P. Tonella, C. Basile, P. Falcarin, M. Torchiano, B. Coppens, B. De Sutter, Understanding the behaviour of hackers while performing attack tasks in a professional setting and in a public challenge, Empirical Software Engineering 24 (2019) 240–286.
[33] P. Falcarin, F. Dainese, Building a cybersecurity knowledge graph with Cybergraph, in: Proc of the 2024 ACM/IEEE 4th International Workshop on Engineering and Cybersecurity of Critical Systems (EnCyCriS) and 2024 IEEE/ACM Second International Workshop on Software Vulnerability, EnCyCriS/SVM '24, ACM, 2024, p. 29–36. doi:10.1145/3643662.3643962.
[34] B. Ma, N. Zhang, G. Liu, L. Li, H. Yuan, Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach, Information Processing & Management 52 (2016).
[35] A. Haghighi, L. Vanderwende, Exploring content models for multi-document summarization, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, NAACL, 2009.
[36] P. Falcarin, P. Chowdhury, E. Carbone, T. Scantamburlo, R. Tripodi, S. Vascon, Legal requirements compliance using NLP and Knowledge Graphs, in: Proc. of the 1st Intern. Workshop on Requirements Engineering for Accountable and Conscious Human-centered AI, REACH-AI 25, IEEE, 2025.