Perspectives

MAX PLANCK LAW

ISSN 2748-3924

Tackling the Cambrian Explosion of Digital Rules with Natural Language Processing

Anselm Küsters
14 March 2023
DOI: 10.17176/20230314-163101-0

ABSTRACT

The recent appearance of numerous rules for the digital sphere is nothing short of a 'Cambrian explosion'. To understand the diverging lines of development, scholars can draw on Natural Language Processing. By revealing the different ways stakeholders use core concepts, these methods can help build a future-proof regulatory framework for the digital age.

1 In 2023, lawyers, legal scholars, and companies will face a host of new laws and regulations as governments worldwide scramble to control the usage and development of novel digital technologies, such as Artificial Intelligence (AI). These efforts are spearheaded by the European Commission, which has launched numerous legislative and non-legislative initiatives in the past years aimed at ensuring a better and fairer distribution of data (Data Governance Act, Data Act), protecting fundamental rights in the digital space (AI Act, EU Digital Principles), and sanctioning anti-competitive practices of American Big Tech companies (Digital Markets Act, Digital Services Act). As the recent gathering of policymakers at the World Economic Forum in Davos made clear, there are also increasing calls for industrial policy measures to promote European cleantech companies and to boost network resilience and cyber security, which might necessitate changes to state aid rules. Some scholars are already seeing a paradigm shift to digital industrial policy in Europe. At the time of writing, the Commission is busy adapting its proposed AI Act to generative AI applications, such as ChatGPT, while also preparing entirely new initiatives targeted at the 'metaverse', the future 3D-enabled digital space that uses virtual reality technology for gaming and business experiences.

2 The regulatory slumber of the past decades, during which the 'Wild West' reputation of the digital space thrived in Europe, has ended. Today, lawyers, entrepreneurs, and consumers in the EU face many new regulations, standards, and norms. This development will even accelerate in the coming years—not least because digitization itself is advancing exponentially. Reading the sociologist Armin Nassehi and his 'Theory of Digital Society', it quickly becomes clear how inevitable digitization actually is and that a mere warning about loss of control is not enough because the new technology is used anyway. This relatively sudden appearance of a vast array of digital rules is nothing short of a 'Cambrian explosion', leading to many diverging and sometimes contradictory lines of development. Both legal scholars and policy-minded observers must consider how to approach such complexity in future.

3 One way to archive the huge corpus of digital rules and to analyse it in a holistic and structured way is to draw on a wide array of methods from Natural Language Processing (NLP). NLP is a branch of computer science dealing with the interaction between computers and human languages that has a wide range of applications in the humanities and social sciences, including the analysis of legal data. An NLP approach assumes that the frequency and position of words in a larger corpus of texts can tell us something about the underlying arguments, topics, and tonality. Legal texts have become increasingly available in digital, machine-readable format, which enables lawyers and legal scholars to use the broad range of NLP tools that have already been successfully employed in other academic fields or industrial use cases. For instance, the IUROPA Project has recently launched a freely available database containing research-ready data on the cases, proceedings, decisions, and judges of the Court of Justice of the EU. Overall, an NLP-centred research perspective advocates using legal documents and accompanying metadata (ie external information about these documents, such as year of publication or name of the presiding judge) as direct objects for quantitative statistical analysis.

4 Using such a 'law as data' approach has several benefits. For instance, the newly gained empirical evidence might be used to test certain legal doctrines—such as the 'equity principle' used by the European Court of Human Rights—to see whether there is a consistent pattern of judicial decision-making. Another option is to use legal data for classification and prediction tasks, not least to improve legal decision-making and to future-proof current tech regulations. Recent attempts use large text corpora of court decisions, eg, from the US Supreme Court or the European Court of Human Rights, to train classifiers to predict the decisions' outcome. Ensuring sufficient explainability of these models is, however, an ongoing challenge. Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch have suggested combining a knowledge-based decision tree with advanced word embeddings like BERT to classify illegal online content automatically, which might support the EU's legal framework against the expression of hatred. Much recent research in the computer science community has also studied the fairness of algorithmic decision-making, often focusing on applications in the legal domain as a case study. For instance, Nina Grgić-Hlača, Elissa Redmiles, Krishna Gummadi, and Adrian Weller have studied the perceived fairness of using different predictive attributes for estimating criminal recidivism risk. A final option is explored in the

remainder of this essay: using NLP methods to reduce complexity in an increasingly turbulent digital age by finding patterns and interesting outliers in legal data.

5 One of the main ways that NLP can help legal scholars in this regard is through text mining and information retrieval. Text mining is a process that involves extracting relevant information from large volumes of text data, such as court cases, disclosed documents, or government reports. For instance, Tilmann Altwicker, Szilvia Altwicker-Hámori, Daniel Gerber, and Anne Peters use hierarchical clustering of legal data to find natural groups and statistical patterns. The method enables them to find co-occurrences of violated legal rules, leading to novel observations. Using text mining techniques such as 'topic modelling', legal scholars can also quickly summarize underlying themes, issues, and opinions in digital regulation, saving valuable time and reducing human errors and cognitive biases that might occur when handling thousands of pages of text. For example, topic modelling-based comparisons between stakeholder contributions, expert reports, and legal reform proposals written as part of the legislative adaptation of EU competition law to the challenges of digitization help trace the transmission of specific ideas and subtle forms of lobbying.

6 Besides clustering and summarizing content from EU documents and regulations, these methods might also help transform the notoriously technical and opaque texts stemming from Brussels into more reader-friendly writings, ultimately increasing transparency, adherence to the law, and democratic accountability. Christian Rauh of WZB Berlin Social Science Center recently used NLP methods to examine around 45,000 EU Commission press releases from 1985 to 2020. His quantitative evidence showed that the Commission typically uses many technical terms comparable to the language common amongst scientists. Evaluated against the press releases of the governments of Great Britain and Ireland, the Commission's sentences are longer, the number of syllables in words is higher, and noun structures occur more frequently. Taken together, this suggests 'technocratic' language. Such a lack of readability and understandability is nowhere more harmful than in the digital context, where politicians might miss a deeper understanding of the very issues they hope to regulate. Patryk Pawlak, who leads the work of the EU Institute for Security Studies on cyber and digital issues, notes that 'the EU needs to build an army of diplomats who can skilfully translate "EU-speak" into a universal and globally acceptable language'. Inspiration might come from projects like the TLDR software, which leverages recent advances in generative AI, namely large language models, to simplify scientific papers into texts that children can easily understand.

7 Besides simplifying language, NLP tools could be used to reflect on the rapidly changing language in the digital age itself. To illustrate the potential, consider the recent study by Daniel Chen, Elliott Ash, and Suresh Naidu, who use a mix of so-called dictionary analysis (a dictionary is basically a long list of unique words that can be used in bag-of-words approaches) and econometrics to show that after attending economics training in the US, participating judges use more economics language and render more conservative verdicts. As a significant disruption of today's economy and society, digitization will affect the law and its

conceptual structure in a similar way to the previous trend towards 'processes of economization' revealed by these authors.

8 In a recent contribution to Max Planck Law 'Perspectives', Marietta Auer conceptualized this shift by looking at key concepts of modern law, such as autonomy, and found that the underlying structure of the digital society might not be so different from the pre-digital one after all. In this context, it is worth pointing out that NLP methods can also be used to identify key concepts in digital regulation and to compare their framing by different actors. In particular, it is possible to leverage Word Embedding Models to quantify evidence of differing understandings of certain concepts. During the current transnational quest to regulate large online platforms like Amazon, researchers have used this technique to investigate whether all stakeholders share the same understanding and use of the relevant terms of the underlying legislative initiatives, such as the Digital Markets Act (DMA) and the Digital Services Act (DSA). They find significant differences in the employment of terms like 'gatekeepers', 'self-preferencing', and 'collusion', raising the question of whether such latent differences in word usage might significantly hinder the consultation process and, ultimately, global enforcement and legal certainty.

9 Another way NLP methods might help legal scholars analyse digital regulation and law-making is through sentiment analysis. Sentiment analysis is a process that involves determining the attitude or emotion expressed in a piece of text. So far, it has been mainly used to analyse subjective statements in Social Media posts such as tweets or consumer reviews on websites like Amazon. Using sentiment analysis, legal scholars can identify the tonality of a particular document, such as whether it sees a specific set of rules as restrictive or permissive. This can be useful for understanding how different stakeholders view a particular regulation and identifying potential areas of controversy or conflicting interpretations. To return to the research mentioned above, which investigated whether all stakeholder groups share the same understanding of the relevant terms in the DMA and DSA, the analysis also included sentiment analysis showing that in some cases, the identified differences in conceptual understanding also came with different attitudes. For example, a concept like 'self-regulatory' was not only used differently by various stakeholders but was also viewed more favourably by medium and large companies and organizations than by small ones. Lawyers and lawmakers might use such evidence to identify supporters and critics of a legislative proposal as well as their main hopes and fears that must be addressed to create consensus.

10 In addition to the approaches mentioned above, legal scholars can use NLP to analyse the form and structure of legal texts, such as lengthy regulations. This is critical in the context of the digital economy, where parallel or overlapping sets of domestic and international rules might apply. This can be done by looking at language sequences, not single words, and using techniques such as the following ones: text reuse, which identifies exact or similar copying of text between documents; parts of speech tagging, which assigns a grammatical tag to each word; named-entity recognition, which tags people, organizations, and geographical locations within texts; as well as dependency parsing, which extracts a directed graph explaining how words are syntactically connected within a sentence. To

begin with, scholars have already experimented with text reuse methods for tracing influence in law-making, for instance, to analyse the extent of parliamentary bill amendments or to identify substantively similar policy proposals in legislation. Using syntax analysis and dependency parsing enables scholars to understand how different clauses and sections of the law are related. This can be useful for understanding how other parts of a regulation fit together and can help legal scholars identify potential ambiguities or inconsistencies. Zooming out to the broader web of digital rules, they can identify similar overlaps and inconsistencies between different sets of rules. Legal scholars can use entity recognition to programmatically determine a regulation's scope and the parties affected by it.

11 Finally, digitized legal data and manual annotations might even contribute to training dedicated algorithms to classify certain parts of regulations or contracts. In other words, the same techniques that necessitate the adaptation of the law in the digital age might also be used to automate specific tasks and thus reduce complexity and burdensome tasks. For example, a team of US-based researchers who focused on automating contract review has recently published a novel dataset of legal clauses from contracts governing company acquisitions. It contains over 47,000 annotations explaining legal clauses extracted from 151 of these contracts and can now be used by legal scholars to train AI systems to review contracts. To achieve similar results for older regulations or contracts typically studied by legal historians, it might be necessary to transform them into machine-readable format through optical character recognition (OCR) software. While outdated fonts such as Fraktur have long made this a time-consuming endeavour, applying new Neural Network-based OCR methods to scan images of books printed between 1487 and 1870 has led to high character and word accuracies. This suggests that diachronic legal datasets, including old texts, can nowadays be constructed quickly and at low cost. In this way, NLP methods can profit from a larger sample of observations while helping to put the digital present and virtual future into a much-needed historical perspective.

12 This essay has considered how NLP can help reduce complexity in a rapidly changing legal landscape, which must adapt to the digital age's evolving reality. In doing so, it has looked at digital technologies both as research objects and as potential methodological tools that might help researchers and regulators deal with new situations. It shares this ambition and approach with the Max Planck Law | Tech | Society (LTS) Initiative, which brings together researchers from across Max Planck Law Institutes and beyond with a particular interest in the latest developments in the field of law and technology. Using the LTS Initiative as a hub for the exchange of ideas on common themes, concerns, and challenges raised by the complex and interdependent relationships between law, technology, and society, members of this initiative—including the author of this essay—frequently discuss the potential of the digital age for making sense of the law (and the other way around). NLP has a role to play in this debate, and there are plenty of examples to learn from. A recent survey estimates that more than 600 papers on legal NLP have been published over the past decade.

13 Natural Language Processing provides a robust set of tools for legal scholars to make sense of the recent 'Cambrian explosion' of digital rules. By using methods such as text mining, sentiment analysis, and dependency parsing, legal scholars can quickly and efficiently identify critical themes, opinions, and entities in digital regulations. This can help them to understand better the scope and implications of the underlying rules as well as to identify potential ambiguities or inconsistencies. NLP can also identify key concepts, their development over time, and the different ways they are used by various stakeholders with diverging interests or traditions of thought. This can assist in the building of a future-proof regulatory framework for the digital age.

SUGGESTED CITATION: Anselm Küsters, 'Tackling the Cambrian Explosion of Digital Rules with Natural Language Processing' *Max Planck Law Perspectives* (14 March 2023), https://law.mpg.de/perspectives/tackling-the-cambrian-explosion-of-digital-rules-with-nlp/, DOI: 10.17176/20230314-163101-0