CaseScope: An Enhanced Search Tool for European Court Cases

Alexandre Correia[1,2], Pedro Evangelista[2], Nádia Soares[2], Eugénio Rocha[3] and Cláudio Teixeira[2]

[1]Department of Physics, University of Aveiro, 3810-193 Aveiro, Portugal
[2]Mindera, Rua Goncalo Cristovão 347 s404, 4000-270 Porto, Portugal
[3]Center of Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, 3810-193 Aveiro, Portugal

Abstract

Natural Language Processing (NLP) is a rapidly growing field of research, enabled by advances in computer power and deep learning models. As a subfield of Artificial Intelligence, NLP can help with tasks such as Named Entity Recognition and Sentiment Analyses by extracting meaningful connections between words from a text. New architectures for neural networks like transformers have been responsible for a great increase in performance at these tasks. These improvements motivated this work, where we look into the extraction of information from legal documents in the CURIA database to develop CaseScope, a search tool that presents users with filters that are machine-generated for court cases from the European Union. Besides enhancing CaseScope's search space with NLP techniques, we also provide a faster way to understand if a case is relevant to the user's search by presenting generated summaries of documents produced with recent models. Main differences between CaseScope and currently available legal search tools are also compared. Features described in this work were developed with a multidisciplinary team, with expertise in many fields, including legal. Throughout our work, we present how CaseScope is built, from data collection to the search interface, to give a better insight into our approach to each step of creating CaseScope.

Keywords

Natural language processing, artificial intelligence, legal search assistance, information extraction, keywords extraction, summarization

1. Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) focused on developing algorithms and models that enable computers to interpret and generate natural language in the form of text or speech. These models can be exploited in a wide range of applications, such as machine translation and sentiment analysis. With recent advancements in NLP, the legal domain is well-placed to take advantage of this progress, since the vast majority of documents produced in court and by legislators are written with specific wording, and the information that can be extracted depends heavily on context.

A deep understanding of the context contained in documents can be the difference between a simple information system that can only see a sequence of words, and a knowledgeable AI, that is capable of advising about a specific legal case.

With the development of Google Transformers [1] and, more recently, OpenAI's GPT [2], many problems that were deemed unsolvable with prior methods became more approachable [3], opening space for new tools to emerge.

In the legal domain, NLP has applications like document analysis, contract review, and legal research. These applications can help lawyers and legal professionals extract insights and information from legal documents, identify patterns in legal language, and automate routine tasks such as contract drafting and review.

Attorneys often face the time-consuming task of searching and reviewing past court decisions relevant to their cases [4]. Existing search tools, though helpful, rely on extensive human effort for the annotation and classification of documents. Moreover, they cannot fully reflect on what the user is looking for, being only able to produce results based on information curated by humans. This dependency on human curation creates a bottleneck, requiring more manpower for handling larger volumes of information. To address this, we explored the use of NLP methods in CaseScope, utilizing the CURIA database, which is a collection of legal cases deliberated by the Court of Justice of the European Union (CJEU) [5]. By extracting information directly from cases, we aim to enhance the search space, present more suitable results, improve information display, and potentially alleviate the workload of human annotators.

With these developments in mind, the present work focuses on the assessment and implementation of relevant techniques to create a search tool for legal documents from the CJEU, using state-of-the-art NLP techniques to address complex problems when searching for informa-

Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal.
alexandredc@ua.pt (A. Correia);
pedro.evangelista@mindera.com (P. Evangelista);
nadia.soares@mindera.com (N. Soares); eugenio@ua.pt (E. Rocha);
claudio.teixeira@mindera.com (C. Teixeira)

tion in large collections of domain-specific documents.

2. State of the art

In the legal domain, NLP tasks that can be considered simple in the open domain, usually become more challenging [6]. Reinforced with the fact that legal experts are necessary to produce annotated datasets, makes the legal domain a very interesting and enticing field to apply state-of-the-art NLP methods.

NLP applied to the legal domain is a subject of study since the beginning of computer science [4]. However, due to the complexity of this domain, it has been difficult to make successful applications. The introduction of transformers has altered this dynamic [7]. Since then, many new models were developed to perform well in this specific domain, achieving state-of-the-art results. Notably, LegalBERT stands out as a leading example of such models, comparing results between adapting a BERT model with further pre-training on domain-specific corpora or pre-training the model from scratch [8].

Existing tools like Casetext [9] and Fastcase [10] already assist lawyers and researchers in their search tasks. Although useful, these tools mainly compute the similarity between documents and searched keywords in order to achieve their result-retrieving capabilities, so they do not yet make use of recent NLP developments as taking advantage of Large Language Models (LLMs) like GPT to enhance their searchable database.

Despite the fact that the number of annotated datasets is slim, the availability of large-scale corpus of legal data, like legal cases and national laws, have been increasing. Programming packages such as the one utilized in this study to retrieve data from the CJEU [11, 12] can simplify the process of accessing such datasets.

3. Methodology

To build an efficient and modern search tool, a variety of factors including the quality of data, storage systems, and the efficiency of the implementation must be considered. To develop a robust and capable tool, it is necessary to carefully consider each of these factors and make informed decisions about how to approach them. In this section, the focus is on the technical details of CaseScope, including data collection, data storage, NLP application, and the search tool itself. A diagram describing how these components are integrated can be analyzed in Figure 1.

The subsection 3.1 covers data collection. A description of methods used to collect the necessary data, the challenges faced during the data collection process and how they were overcome is presented. In 3.2, we discuss data storage, which systems were used to store the collected data and explain the rationale behind the choices made. The 3.3 and 3.4 subsections focus on the NLP application and on the search tool itself. Algorithms used to enhance the data are described and the user interface of CaseScope and how it was designed to facilitate the user's interactions with the system is explained.

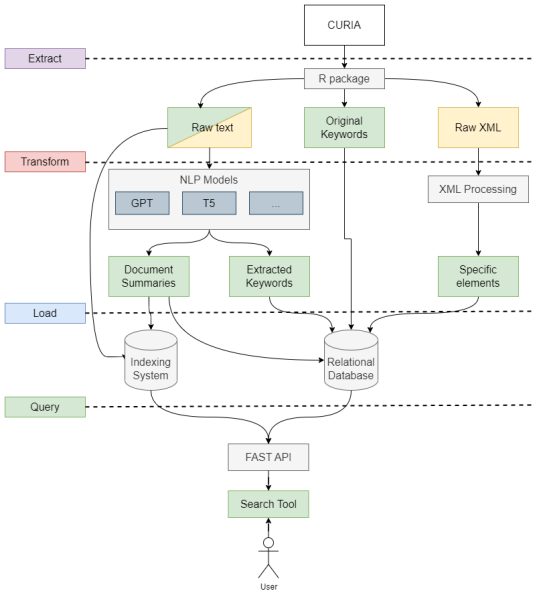

Figure 1: Diagram of how data flows in CaseScope. After extraction, data is transformed through different preprocessing methods. Then we load the transformed data into the proper storage systems and make it available to be queried by users.

3.1. Data collection

As stated in the introduction, CaseScope has the objective of enabling enhanced search capabilities on documents published by the CJEU. This collection of documents is also known as CURIA. To collect all available data we opted to utilize a pre-existing package for the programming language R capable of executing queries to a bigger database called Eur-Lex [13], where CURIA documents are also present.

The eurlex R-package was created with this exact objective in mind, to give a cleaner option of data retrieval to political scientists and legal scholars working with data from the EU. To achieve this, an interface was created with methods written in R that enable researchers to retrieve data from Eur-lex, without having to learn how to compile SPARQL queries. Through this package is possible to request certain metadata, execute pre-made or any other manually input query, as well as download XML files from court cases [11]. Utilizing this tool, we were able to collect more document data than just text. Information was retrieved in three distinct ways.

3.1.1. XML file

XML files have a tree-shaped representation of data, making them ideal for organizing and storing metadata. Additionally, having a well-established category hierarchy enables automation in both storage and search processes.

We utilized these characteristics to establish an automated retrieval system for relevant data, which forms the basis for our filtering system in searches. The retrieved data includes identifiers, dates, document type, court and jury information, subject matters, treaties, and the document's link. Additionally, we collected a set of concepts discussed in each case. Both subject matters and concepts provide a high-level representation of the document's topics. Subject matters offer insight into the main subject, while concepts are organized in a hierarchical structure, associating each document with one or more concepts.

3.1.2. Keywords

The document's title, retrieved using the R-package described before, is a header that provides the name of the two main parties involved, the identifier of the case, and associated keywords. These keywords are what we were looking for, since the first two could already be collected from the XML file.

Keywords found in this section are written without hard rules. There is no pre-determined set of keywords that can be applied to a document like there is for subject matters and concepts. This means more specific context can be given to each case, but it also means it is harder to create connections between documents based on these keywords.

3.1.3. Legal text

The text of each document was also retrieved. Text is the primary form of data that can be processed with NLP techniques. From legal texts, it is possible to extract information such as parties involved, keywords, and citations.

The structure of CURIA documents is generally consistent and follows a standard format, which includes sections such as the introduction, legal context, arguments of the parties, assessment of these arguments, and decision. This structure helps to ensure that all relevant information is included in the document and makes it easier for readers to understand the legal issues at stake.

The introduction section of these documents provides an overview of the case and the issues at stake. The legal context section enumerates all the relevant EU law and any national law or international treaties that are relevant to the case. This section is critical as it helps readers to understand the legal framework within which the case was heard. The arguments section outlines the parties' arguments, which is important for understanding the legal issues being deliberated. The assessment section is the most important section of the document, as it is where the court analyzes the legal and factual issues and applies the relevant law to the facts of the case. The decision section set out the court's final decision and any specific orders or directions made by the court.

Legal documents are often written with highly technical and specialized language, which can make it very difficult to comprehend and extract the correct meaning. They can also be extremely lengthy and numerous, so summarized legal documents and identification of similar subjects in different texts are also useful. By utilizing NLP techniques in these texts we hoped to ease that hardship, helping lawyers, legal researchers, and other professionals analyze legal documents more effectively.

3.2. Data storage

Since we had two different types of data, structured metadata, and unstructured textual data from documents, different methods of storage were utilized for each type of data. For information that had a clear structure, a relational database was designed and developed to provide organized and searchable storage for data, including case details, enabling efficient filtering and retrieval, with related tables and citation tracking. For textual data, like the full text of each document, Elasticsearch [14] was utilized as our indexing system, enabling scalable storage and retrieval of large volumes of text with complex queries, suitable for searching through the full text of Curia documents.

3.3. NLP application

By incorporating NLP capabilities into a legal search tool, users can easily and quickly find the information they need, saving them time and improving the accuracy of their research.

In CaseScope, we utilized NLP models to accomplish two tasks. The first was enhancing the keyword field for each document. This means taking advantage of NLP capabilities to understand word dependencies and importance in a text, and from that retrieve new keywords that can be associated with the document. By doing this we were able to expand beyond the human-generated keywords, making it easier to find documents that are related to the user search. The second task where we applied NLP models was document summarization. These automatically generated summaries of documents can save users a significant amount of time when reviewing large volumes of documents. By displaying them together with each respective document in a list of search results in CaseScope's interface, the user can better understand which documents they should review first.

To accomplish these two tasks, two different models were tested and their results were stored. Firstly we made

```
{
    "summary": "The Supreme Court of Cassation in Italy requested a preliminary ruling (…)",
    "model": "gpt-3.5-turbo",
    "max_tokens_input": 1500,
    "max_tokens_output": 300,
    "splited": true
},
{
    "summary": "The Court of Justice of the European Union ruled that Article 49 TFEU (…)",
    "model": "gpt-4",
    "max_tokens_input": 5500,
    "max_tokens_output": 500,
    "splited": false
},
{
    "summary": "A preliminary ruling was made by the corte suprema Di cassazione (…)",
    "model": "t5",
    "max_tokens_input": 1000,
    "max_tokens_output": 100,
    "splited": true
}
```


Figure 2: Example of JSON file where are stored summaries generated with different models for the same document. These are then imported into the relational database.

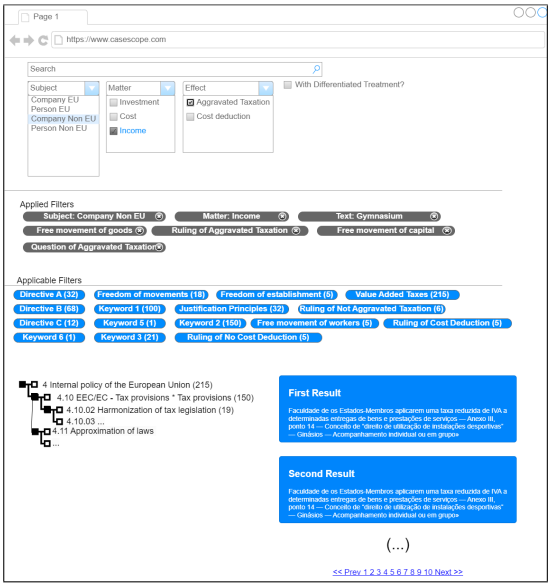

Figure 3: Mockup design for CaseScope's interface. This possible design for a specific version directed only at fiscal search. On the top is the free text search, followed by three dropdown menus with main keywords, where the user can select from the options. Next, there are other keywords, which the user can select and unselect as they see fit. In the bottom-left corner, there is a representation of the concept tree intended to present the user with a complementary way of searching through the results. The blue rectangles on the bottom-right corner represent a possible display of the results list.

use of the T5 model [15]. This is a model released in 2019, that can run on a local machine, which made it a great choice to begin with. Besides the T5 capabilities, the model is somewhat outdated in the NLP ever-evolving field. So to try to achieve the best results possible, we then utilized models from the GPT family [16], the "gpt-3.5-turbo" and the "gpt-4" [2], through OpenAI's API. The first is a more affordable, equally capable version of "gpt-3.5" and the latter is, at the moment, the most capable model from the GPT family.

The results produced by these models were stored in a JSON format as shown in Figure 2. They were then placed in the relational database, as fields in the "document" table, one for generated keywords, and another for generated summaries.

3.4. Search tool

One of the objectives of this work was the creation of a competent legal search tool that can distance itself from existing tools by taking advantage of the most recent developments in the NLP field. Legal search tools can be used in a variety of legal contexts. For example, legal professionals may use legal search tools to find relevant case law when preparing legal briefs or arguing cases in court, and researchers may use legal search tools to identify trends in legal decisions or explore the evolution of legal concepts over time.

We developed CaseScope with a "search first - filter after" approach. In CaseScope, users can search for specific concepts and then apply filters to the results interactively. Figure 3 shows an early proposed interface, initially designed for searching fiscal documents but scalable to other domains. The interface features a search bar for text queries and three dropdown menus for applying filters. Users can apply multiple filters and see the number of cases resulting from each filter. After each search or filter application, results are displayed with their corresponding summaries and main keywords.

This is the primary way of searching with CaseScope but an advanced search interface can also be present to let the user search freely with filters without having to go through an initial search.

CaseScope, through the use of its API, may also be integrated with any interface, and even exploited as an extension for existing search tools.

4. Results discussion

The use of a pre-existing package for the programming language R allowed for the execution of queries to the Eur-Lex database, enabling the collection of more data than just text from the documents. In addition, this package provided a cleaner option for data retrieval, as it was not necessary to learn how to compile SPARQL queries.

One of the challenges we faced while collecting data was keyword retrieval from the documents' titles with

```
"x"
"1" "Judgment of the Court (Eighth Chamber) of 6 October 2022.#Agenzia delle
Entrate v Contship Italia SpA.#Requests for a preliminary ruling from the Corte
suprema di cassazione.#References for a preliminary ruling – Direct taxation –
Freedom of establishment – Corporate income tax – Measures to prevent tax avoidance
by shell companies – Determination of taxable income on the basis of presumed
minimum income – Exclusion from the scope of those measures of companies and
entities listed on national regulated markets.#Joined Cases C-433/21 and C-434/21."
```

Figure 4: Example of a title retrieved from a document using the R-package.

text separation techniques. Figure 4 shows an example of a document title. From that title, our objective was to retrieve the following keywords: "References for a preliminary ruling", "Direct taxation", "Freedom of establishment", "Corporate income tax", "Measures to prevent tax avoidance by shell companies", "Determination of taxable income on the basis of presumed minimum income", "Exclusion from the scope of those measures of companies and entities listed on national regulated markets".

Although seemingly simple, this task became challenging due to different documents using different characters to separate keywords, as well as differences in spacing. Those challenges were eventually overcome by understanding every difference and taking measures to correct each of them, but this means regular reviews are still necessary in case new documents are written differently, making the process not fully autonomous.

Another task we aimed at completing was the collection of the questions referred to the court. These questions are an important section of Curia documents, and from them, it is possible to understand what is being argued in the case, and if the court responded or not to the questions referred. For this task results were still incoherent due to inconsistent document formatting, making the retrieval of these questions directly from text a very difficult task to automate.

From the XML files, retrieval of information was simpler. This was our main source of information to create filters, enabling the collection of basic information on each case, such as parties involved, identifiers, and citations. It also permitted the use of the already well-established tree of concepts curated by the EU to categorize cases.

With the application of NLP models, it was possible to produce summaries for documents. The presentation of accurate and reliable summaries for large documents has great potential to improve the productivity of legal search work. If a well-written summary of a document is available while the user is only browsing through the results they can rapidly understand if a given document is relevant or not for their work, accelerating this discovery process and making the pool of documents that have to be fully analyzed smaller.

To produce these summaries, we tested T5, "gpt-3.5-turbo," and "gpt-4" models and conducted a qualitative evaluation. Summaries generated by the T5 model were either lengthy or missed the main points of the case, often including irrelevant information such as the decision on costs. Moving on to "gpt-3.5-turbo," we observed improvements with more concise phrases and relevant information. Although artifacts like the inclusion of the original language of the case were still present, the model showed promise. Lastly, we tested the "gpt-4" model, which offers the advantage of handling larger texts without the need for chunking. The results are promising, as demonstrated by the example of a summary generated using this model:

"The Court of Justice of the European Union ruled that Article 49 TFEU does not preclude national legislation restricting the ground for exclusion from the scope of measures to prevent tax avoidance by shell companies to companies whose securities are traded on national regulated markets. This exclusion does not apply to other companies, whether national or foreign, whose securities are not traded on national regulated markets but are controlled by companies and entities listed on foreign regulated markets."

Besides summary generation, NLP models were also applied to enhance CaseScope's search space. With the help of GPT models, more keywords were collected for each document. Taking the document with its title presented in Figure 4 as an example, in addition to the keywords already listed before, we were then capable of collecting the following keywords by utilizing "gpt-4": "Italian laws", "tax avoidance measures", "securities", "national regulated markets", "freedom of establishment", "discrimination", "national companies", "foreign companies", "parent company", "shell companies", "European Union law".

As we can see the existence of overlapping keywords indicate successful extraction of useful keywords by the model. The generated keywords are smaller in size than the human-labeled ones, which helps in expanding the search space with a broader and more general group of keywords. In some cases, the model separates a single keyword from the human-labeled set into two distinct keywords. For instance, the keyword "Measures to prevent tax avoidance by shell companies" is divided into "tax avoidance measures" and "shell companies" in the set of generated keywords.

With the objective of better understanding how CaseScope relates to legal search tools available on the market, we described the features of four well-established and reliable legal search tools together with CaseScope's features. This description can be reviewed in Table 1.

For CaseScope, we marked features that are in our roadmap, which doesn't exclude the addition of new features in a later review. AI/Machine Learning will at least be present in the form of what was discussed in this work, the enhancement of the search field with the assistance

Table 1
Feature description of CaseScope and some of the most prominent legal search tools (based on https://www.capterra.com/legal-research-software/).

Feature	CaseFleet [17]	Casetext [9]	Fastcase [10]	LiquidText [18]	CaseScope
AI/Machine Learning	✓	✓	✓	○	✓
Annotations	✓	✓	○	✓	○
Brief Analytics	○	✓	○	✓	○
Case Alerts	✓	✓	✓	○	○
Case Law Research	○	✓	✓	✓	○
Change Tracking	○	✓	○	✓	○
Data Visualization	✓	✓	✓	✓	✓
Query Suggestions	○	○	✓	○	○
Search History	○	✓	✓	✓	✓
Search/Filter	✓	✓	✓	✓	✓
Self-Service Search	✓	✓	✓	○	○
Statutes Research	○	✓	✓	✓	○
Tax Law Research	○	○	○	✓	○
Third-Party Analysis Integration	○	○	○	✓	○
Summarization	○	○	○	○	✓
Keyword Extration	○	○	○	○	✓

of NLP techniques. Data visualization and Search/Filter are core concepts of a search tool, so those are mandatory. Search History is a feature we understand to be essential to have a good workflow, enabling the user to pause a search session and resume it at a later time.

The work here described shows our focus and ability to develop other features that are not yet available in previously discussed tools, and that is the differentiator factor for CaseScope, enhancing search capabilities and the workflow of legal practitioners and researchers. CaseScope can be integrated with any interface through its API, and could even be leveraged by existing legal search tools to extend their capabilities, since it is a robust application developed with insights from legal practitioners.

5. Conclusion

The main objective of this work was to review the various necessary methods to develop a search tool for CURIA legal documents and apply them. Advanced NLP models were utilized to tackle the intricate challenges involved in searching for information in vast collections of domain-specific documents.

We first looked into how data from the CURIA documents could be retrieved to feed CaseScope, through an R-package named "eurlex" and further preprocessing. Then we went over our approach to storing this said data, by utilizing a relational database and an indexing system, for structured and unstructured data, respectively.

With the application of these NLP models, we were able to produce summaries for documents and collect more keywords for each document, improving CaseScope's search space. Overall, the results show the potential of NLP models in legal research and document analysis, and the benefits they can bring to legal practitioners and researchers.

Future work for CaseScope can focus on evaluating the results of NLP models systematically using automated metrics for summaries evaluation, enhancing confidence and identifying areas for improvement. Additionally, implementing a Q&A system to retrieve answers to specific legal questions and integrating them into CaseScope's searchable database would be valuable. Furthermore, ensuring an extensible tool that stays updated with the latest CURIA documents requires creating a pipeline to automate the methods described in this work, ensuring a stable and useful legal search tool.

Acknowledgments

This work is part of a joint venture project where we worked with Morais Leitão, Galvão Teles, Soares da Silva & Associados, who as lawyers and domain experts approached us with this innovative project to be developed combining legal and technical expertise. We would like to express our gratitude to Cláudia Baptista, Ana Pedro Castro, Carlos Coelho and António Queiroz Martins for their contribution to this research. The fourth author was partially supported by the Center for Research and Development in Mathematics and Applications (CIDMA), through the Portuguese Foundation for Science and Technology, reference UIDB/04106/2020 and UIDP/04106/2020.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[2] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[3] L. Tunstall, L. Von Werra, T. Wolf, Natural language processing with transformers, " O'Reilly Media, Inc.", 2022.

[4] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How does nlp benefit legal system: A summary of legal artificial intelligence, 2020. arXiv:2004.12158.

[5] The court of justice of the european union, 2023. URL: https://curia.europa.eu/jcms/jcms/j_6/en/, accessed on April 17, 2023.

[6] J. Savelka, V. R. Walker, M. Grabmair, K. D. Ashley, Sentence boundary detection in adjudicatory decisions in the united states, Traitement Automatique Des Langues (2017) 21–45. doi:10.47164/ijngc.v14i1.1004.

[7] A. Gillioz, J. Casas, E. Mugellini, O. A. Khaled, Overview of the transformer-based models for nlp tasks, in: 2020 15th Conference on Computer Science and Information Systems (FedCSIS), 2020, pp. 179–183. doi:10.15439/2020F20.

[8] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, 2020. arXiv:2010.02559.

[9] Casetext, 2023. URL: https://casetext.com/, accessed on April 30, 2023.

[10] Fastcase, 2023. URL: https://www.fastcase.com/, accessed on April 30, 2023.

[11] The court of justice of the european union, 2023. URL: https://cran.r-project.org/web/packages/eurlex/vignettes/eurlexpkg.html, accessed on April 17, 2023.

[12] M. Ovádek, Facilitating access to data on european union laws, Political Research Exchange 3 (2021) 1870150. URL: https://doi.org/10.1080/2474736X.2020.1870150. doi:10.1080/2474736X.2020.1870150.

[13] Eur-lex: Access to european union law, 2023. URL: https://eur-lex.europa.eu/homepage.html?locale=en, accessed on April 13, 2023.

[14] Elasticsearch, 2023. URL: https://www.elastic.co/, accessed on April 30, 2023.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[16] A. Radford, K. Narasimhan, Improving language understanding by generative pre-training, 2018.

[17] Casefleet, 2023. URL: https://www.casefleet.com/, accessed on April 30, 2023.

[18] Liquidtext, 2023. URL: https://www.liquidtext.net/, accessed on April 30, 2023.