Legal Knowledge and Information Systems
J. Savelka et al. (Eds.)

doi:10.3233/FAIA241262

Combining Rule-Based and Machine Learning Methods for Efficient Information Extraction from Enforcement Decisions

Harry NAN [a,1], Maarten MARX [b] and Johan WOLSWINKEL [a]
[a] *Tilburg University, Tilburg Law School*
[b] *University of Amsterdam, Faculty of Science*
ORCiD ID: Harry Nan https://orcid.org/0009-0003-8291-8818, Maarten Marx https://orcid.org/0000-0003-3255-3729, Johan Wolswinkel https://orcid.org/0000-0001-8404-5027

Abstract. This paper presents an effective and efficient approach for automatic extraction of key features from enforcement decisions, such as their legal basis and their legal effect, by strategically applying a Large Language Model (LLM) on top of rule-based methods. Initially, rule-based methods identify candidate sentences within these decisions containing these features, after which these sentences are analyzed by GPT-3.5 to extract the features. This approach is efficient as it reduces the input and number of resources needed for effective and context aware information extraction. Furthermore, other features that have not been subject to a rule-based selection first can be extracted by an LLM from the same set of candidate sentences when they exist in close proximity of each other.

Keywords. information extraction, enforcement decisions, rule-based methods, machine learning methods, large language model, text segmentation

1. Introduction

With the increasing public availability of legal documents, appropriate metadata are of utmost importance to facilitate a smooth processing and re-use thereof [1]. For types of legal documents that do not have a well-established tradition of public disclosure, such as administrative decisions, uniform standards are lacking. This makes it necessary to extract key information from the texts of these documents in both an effective (i.e. accurate) and an efficient manner [2]. However, Information Extraction (IE) on these documents can be highly challenging, primarily due to their considerable length and unstructured nature [3]. Whereas rule-based techniques are often ineffective when dealing with varied patterns [4], machine learning techniques, in particular Large Language Models (LLMs) such as Generative Pre-Trained Transformers (GPT) suffer from the inability to

[1]Corresponding Author: Harry Nan, h.nan@tilburguniversity.edu. This research is part of the research project CITaDOG from Tilburg University.

process large texts [5,6]. In the legal field, these techniques have mainly been researched in isolation, focusing on improving rule-based systems (e.g. [4]) or on enhancing the contextual capabilities of machine learning models (e.g. [7]), by splitting input into multiple prompts for IE. This paper proposes a combination of rule-based methods with a LLM by selecting sentences from enforcement decisions that contain feature indications with the use of named entity recognition (NER) and regular expressions. Subsequently, these sentences are processed by GPT-3.5 to achieve effective information extraction, i.e. extraction of the correct key features. This approach is considered to be efficient as it first applies rule-based methods in order to minimize input text for the LLM. This paper explores the effectiveness of this hybrid approach with the following research question:

***RQ**: To what extent can GPT-3.5 in combination with rule-based sentence selection be effective for information extraction from enforcement decisions?*

2. Background

Legal information extraction differs from other domains due to the length of the documents, their complex structures, and the jargon used [8]. In earlier years, rule-based methods and tasks like regular expressions, parsing, and NER were mainly used to extract information from legal data [9,10] next to being applied to related legal tasks (e.g., classification and semantic network creation) [2]. These methods excel when data consist of clear patterns, like laws and references [4,11,12], but may struggle with flexibility and context-dependent tasks [12].

More recent studies show that generative LLMs show promising results for information extraction with minimal resources [7], being applied to various legal data (e.g. [2,6]) and tasks (e.g. [13,14]). Zero-shot learning has demonstrated near-human accuracy without extensive training data, making these models effective in resource-constrained environments [15]. However, certain challenges remain: poor performance (or ability) in processing large text volumes [6], uncertainties and limitations in specific languages (e.g. [16]), the generation of noise and hallucinations, and the importance of prompt engineering [5] have been highlighted.

Our work combines the strengths of both approaches. By using rule-based methods to reduce input text for LLMs, we address the token limitations of such models like GPT-3.5 [6], while allowing them to leverage context-aware capabilities [17].

3. Methodology

We have selected enforcement decisions of two Dutch administrative bodies, namely the Kansspelautoriteit (Dutch gambling authority, KSA)[2] and the Autoriteit Financiële Markten (Dutch financial markets authority, AFM)[3]. The data are publicly available on their websites, but for this research obtained from the platform Woogle.[4] We consider two types of enforcement decisions: administrative fines and administrative penalties.

[2] https://kansspelautoriteit.nl/
[3] https://afm.nl/
[4] https://woogle.wooverheid.nl; downloaded in April 2024.

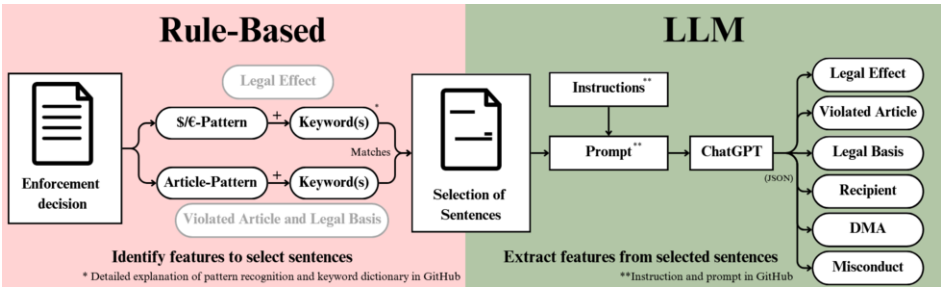

Figure 1. Pipeline of methodology.

Whereas administrative fines impose an *unconditional* obligation to pay, administrative penalties only impose a *conditional* obligation to pay if non-compliance is not terminated within a certain time period. We identify six key features that every enforcement decision should contain based on Dutch law: a decision-making authority (DMA), a recipient, a legal basis, a legal effect, the actual misconduct, and the violated legal provision. The selected data included 299 enforcement decisions: 175 administrative fines (KSA: 65, AFM: 110) and 124 administrative penalties (KSA: 55, AFM, 69). The GitHub-page[5] shows a deeper analysis of this data.

To answer the research question, we follow a two-step approach as shown in Figure 1. Firstly, since rule-based methods excel in the extraction of information when patterns are clear and structures are similar, we identify three key features of administrative decisions that seem to consist of such identifiable patterns or structures, namely Legal Effect, Violated Article, and Legal Basis. Regular expressions and SpaCy's[6] NER and Part of Speech (POS) techniques are used to detect patterns in money references and legal provisions, which are combined with keyword matching to identify these features,[5] after which the sentence identified and its direct neighbors are selected with removal of any content overlap. This pipeline and techniques are combined, as similar approaches on legal data have proven effective and the key features align well with SpaCy's pipeline [4,11].

Secondly, the LLM, more specifically gpt-3.5-turbo-0125,[7] analyses these selected sentences and extracts the key features in a context-aware and zero-shot manner. GPT-3.5 is used as it has been widely adopted in the legal field with promising results and allows for contextually accurate extraction in a resource-efficient manner [17].

In addition, qualitative analysis suggests that the other three features that might lack identifiable patterns (recipient, DMA, misconduct) can be found within the same set of candidate sentences (Figure 2). It is therefore expected that a collection of (neighboring) candidate sentences is sufficient to extract all key features of administrative decisions. All six features of enforcement decisions are therefore included in the prompt. The full prompt presented in the GitHub page[5] includes general instructions explaining the context and the task, followed by an explanation of each feature and how to extract this feature, and an instruction to provide output in JSON-format. The prompt is slightly adjusted between the two types of frameworks to account for their different legal implications.

[5] https://github.com/Harry-Nan/IE-administrative-decisions
[6] SpaCy's pre-trained pipeline 'nl_core_news_lg'.
[7] https://platform.openai.com/docs/models/gpt-3-5-turbo

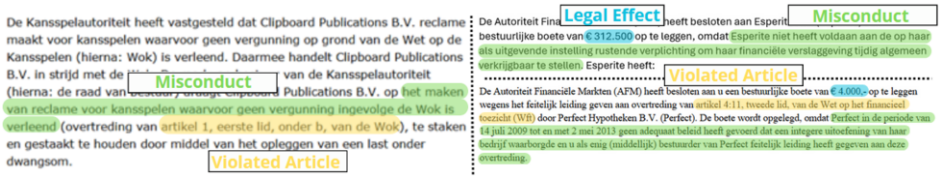

Figure 2. Example of qualitative analysis where features without identifiable patterns (in green) are often found in close proximity to features with identifiable patterns.

Table 1. Percentage Agreement (PA) for each information type to assess inter-rater reliability.[5]

	Legal Effect	Violated Article	Legal Basis	Recipient	DMA	Misconduct
PA	91.7%	100%	76.9%	83.3%	100%	90.9%

The pipeline is evaluated using precision, recall, and F1-score, based on the LLM's output. A golden standard is created from 40 hand-annotated decisions following an annotation protocol supervised by a legal expert, which contains 10 decisions for each combination of enforcement decision type and administrative body (KSA/AFM).[5] Inter-rater reliability is measured with percentage agreement scores between two annotators,[8] by dividing the amount of agreements by the sum of agreements and disagreements. This is displayed in Table 1, showing high agreement scores. In cases of disagreement, the legal domain expert made the final decision for the golden standard.

Macro-averaged precision is calculated as the ratio of correctly extracted information to the total amount of extracted information per document, averaged across key features. For features that can be long and have multiple correct notations (Recipient, Misconduct, and DMA), Bilingual Evaluation Understudy (BLEU) scores are calculated for precision. BLEU evaluates model-generated text whilst allowing for multiple correct notations [18], and is therefore chosen as the method to evaluate these features (n=1). Macro-averaged recall is calculated by dividing the amount of correctly extracted information by the amount of information extracted from the golden standard. For similar reasons as stated above, Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1) is being used to calculate the recall for Recipient, Misconduct and DMA [18]. The F1-score is calculated as a combined measure of precision and recall.

4. Results & Discussion

Table 2 shows the evaluation of our hybrid model. Features with identifiable patterns are extracted accurately, with high F1-scores, but lack consistency in scores across categories of enforcement decisions and administrative bodies. High scores suggest that the rule-based approach correctly identified the feature and that the LLM correctly extracted this feature from the selected sentences. However, inconsistencies, such as with Legal Basis and Violated Article, may be due to diverse wording or overly compact sentences that hinder LLM comprehension. Future research should focus on using more sophisticated approaches by refining rule-based methods to handle diverse wording and optimize sentence selection (e.g. [4]), ensuring correct feature identification and improving the

[8] Two Master students with experience in reading administrative decisions.

Table 2. Precision, Recall, and F1-scores for the two governmental bodies Kansspelautoriteit (KSA) and Autoriteit Financiële Markten (AFM) for the six information types as defined in section 3.

		Administrative Body					
		KSA			AFM		
Category	**Feature**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Fines	Legal Effect	0.950	1.000	0.974	1.000	1.000	1.000
	Violated Article	1.000	1.000	1.000	1.000	0.900	0.947
	Legal Basis	1.000	1.000	1.000	0.222	0.222	0.222
	Recipient	0.800	0.833	0.816	1.000	1.000	1.000
	DMA	0.874	0.920	0.896	1.000	1.000	1.000
	Misconduct	0.811	0.818	0.814	1.000	1.000	1.000
Penalties	Legal Effect	0.750	0.789	0.769	0.625	0.676	0.650
	Violated Article	0.700	0.737	0.718	0.658	0.738	0.696
	Legal Basis	1.000	1.000	1.000	0.300	0.300	0.300
	Recipient	0.818	0.825	0.821	0.797	0.800	0.798
	DMA	0.801	0.840	0.820	0.900	0.900	0.900
	Misconduct	0.954	0.958	0.956	0.659	0.724	0.690

LLM's ability to interpret context across different decision types. Features like Recipient, DMA, and Misconduct, which lack identifiable patterns, show more consistent precision and recall scores, suggesting that the extracted sentences from other features mostly included these features, though the LLM occasionally introduced noise or hallucinations. The recall often being higher than precision hints at some extraction inaccuracies.

The results indicate that a hybrid approach can be effective for extracting not only easily identifiable features, but also less easily identifiable features in close proximity with these former features, while reducing the LLM's input, thereby also increasing efficiency. However, limitations remain, as rule-based methods require flexible patterns, and LLMs need sufficient context to extract features accurately. Since this research only demonstrates the effectiveness of a specific hybrid approach but does not compare its performance to other approaches, future research should refine these approaches, compare their relative effectiveness, and test them on different types of decisions and LLMs to ensure generalizability.

5. Conclusion

This paper combines rule-based methods with LLMs for information extraction from Dutch enforcement decisions. Rule-based methods identify candidate sentences with consistent patterns, reducing the input for the LLM. This allows the LLM to extract features from more relevant text within those decisions, thereby improving efficiency and context-awareness. This study shows that features without clear patterns, which are very challenging for rule-based methods, can be effectively extracted by LLMs when applied to candidate sentences of more identifiable features if they are in close proximity to each other. Future research should explore more sophisticated rule-based methods to ensure generalizability with regard to other administrative decisions to handle diverse wording and optimize sentence selection for the LLM. Nonetheless, these results suggest that a hybrid approach can be both efficient and effective, providing a foundation for large-scale analysis of administrative decisions.

References

[1] Sansone C, Sperlí G. Legal information retrieval systems: State-of-the-art and open issues. Information Systems. 2022;106:101967. doi: 10.1016/j.is.2021.101967.

[2] Giri R, Porwal Y, Shukla V, Chadha P, Kaushal R. Approaches for information retrieval in legal documents. In: 2017 Tenth International Conference on Contemporary Computing (IC3). IEEE; 2017. p. 1-6. doi: 10.1109/IC3.2017.8284324.

[3] Oard DW, Baron JR, Hedin B, Lewis DD, Tomlinson S. Evaluation of information retrieval for E-discovery. Artificial Intelligence and Law. 2010;18:347-86. doi: 10.1007/s10506-010-9093-9.

[4] van Opijnen M, Verwer N, Meijer J. Beyond the experiment: the eXtendable legal link eXtractor. In: Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL); 2015. Available at SSRN: https://ssrn.com/abstract=2626521.

[5] Zhang J, Chen Y, Liu C, Niu N, Wang Y. Empirical evaluation of ChatGPT on requirements information retrieval under zero-shot setting. In: 2023 International Conference on Intelligent Computing and Next Generation Networks (ICNGN). IEEE; 2023. p. 1-6. doi: 10.1109/ICNGN59831.2023.10396810.

[6] Zin MM, Nguyen HT, Satoh K, Sugawara S, Nishino F. Information Extraction from Lengthy Legal Contracts: Leveraging Query-Based Summarization and GPT-3.5. In: Legal Knowledge and Information Systems. IOS Press; 2023. p. 177-86. doi: 10.3233/FAIA230963.

[7] Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. International Journal of Medical Informatics. 2024;183:105321. doi: 10.1016/j.ijmedinf.2023.105321.

[8] Zadgaonkar AV, Agrawal AJ. An overview of information extraction techniques for legal document analysis and processing. International Journal of Electrical & Computer Engineering (2088-8708). 2021;11(6). doi: 10.11591/ijece.v11i6.pp5450-5457.

[9] Jackson P, Al-Kofahi K, Tyrrell A, Vachher A. Information extraction from case law and retrieval of prior cases. Artificial Intelligence. 2003;150(1-2):239-90. doi: 10.1016/S0004-3702(03)00106-1.

[10] Leitner E, Rehm G, Moreno-Schneider J. Fine-grained named entity recognition in legal documents. In: International Conference on Semantic Systems. Springer; 2019. p. 272-87. doi: 10.1007/978-3-030-33220-4$_2$0.

[11] Peikert S, Birle C, Al Qundus J, Le Duyen Sandra V, Paschke A. Extracting References from German Legal Texts Using Named Entity Recognition. In: Legal Knowledge and Information Systems. IOS Press; 2022. p. 231-6. doi: 10.3233/FAIA220472.

[12] Varga D, Gojdič M, Szoplák Z, Gurskỳ P, Horvát S, Krajci S, et al. Extraction of Legal References from Court Decisions. In: Information Technologies – Applications and Theory 2023. vol. 3498; 2023. p. 89-95. Available at: ceur-ws.org.

[13] Gray M, Savelka J, Oliver W, Ashley K. Can GPT Alleviate the Burden of Annotation? In: Legal Knowledge and Information Systems. vol. 379. IOS Press; 2023. p. 157-66. doi: 10.3233/FAIA230961.

[14] Belfathi A, Hernandez N, Monceaux L. Harnessing GPT-3.5-turbo for Rhetorical Role Prediction in Legal Cases. In: Legal Knowledge and Information Systems. vol. 379; 2023. p. 187-96. doi: 10.3233/FAIA230964.

[15] Wei X, Cui X, Cheng N, Wang X, Zhang X, Huang S, et al. Zero-shot information extraction via chatting with ChatGPT. arXiv. 2023;2302.10205. doi: 10.48550/arXiv.2302.10205.

[16] Tong B, Chengzhi Z. Extracting Chinese Information with ChatGPT: An Empirical Study by Three Typical Tasks. Data Analysis and Knowledge Discovery. 2023;7(9):1-11. doi: 10.11925/infotech.2096-3467.2023.0473. Translated to English.

[17] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023;3:121-54. doi: 10.1016/j.iotcps.2023.04.003.

[18] Yang A, Liu K, Liu J, Lyu Y, Li S. Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In: Proceedings of the Workshop on Machine Reading for Question Answering; 2018. p. 98-104. doi: 10.18653/v1/W18-2611.