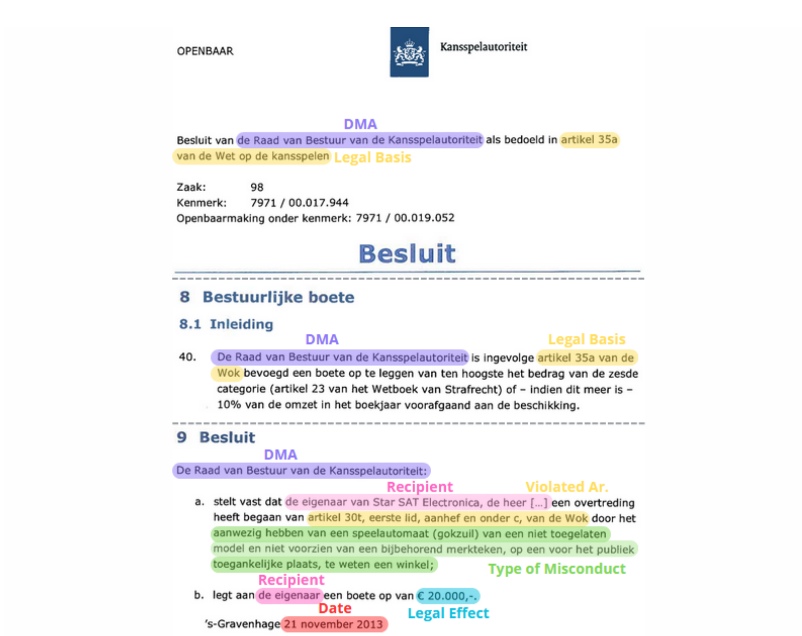
Combining rule-based and machine learning methods for efficient information extraction on administrative decisions

submitted in partial fulfillment for the degree of master of science

Harry Nan
13460854

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 29.06.2024



	UvA Supervisor	External Supervisor
Title, Name	dr. M.J. Marx	prof.mr.dr. Johan Wolswinkel
Affiliation	Faculty of Science - Informatics Institute	Tilburg Law School
Email	M.J.Marx@uva.nl	C.J.Wolswinkel@tilburguniversity.edu

ABSTRACT

This project applies a combination of rule-based methods with machine learning methods to achieve efficient information extraction from large bodies of text, more specifically Dutch administrative decisions. This is done by using rule-based techniques to identify key sentences that contain information to be extracted and are analyzed and extracted by a large language model, ChatGPT. Different types of information can be extracted this way, irrespective of clearly identifiable patterns or structures. The results show that the overall information extraction process is effective, but is dependent on the flexibility and ability of rule-based methods to correctly identify types of information, and an effective sentence extraction with sufficient information for the machine learning method to accurately shape the context. The project highlights the need for a thorough analysis of the information to be extracted and its context within the data to understand what approach is needed for efficient and accurate information extraction.

KEYWORDS

information extraction, rule-based methods, machine learning methods, legal data

GITHUB REPOSITORY

https://github.com/Harry-Nan/IE-administrative-decisions

1 INTRODUCTION

In the current landscape of data-driven decision making in the legal field, the ability to efficiently and accurately extract meaningful information from various unstructured texts is of great importance [7]. Traditional rule-based approaches, such as Regular Expressions (RegEx) and Named Entity Recognition (NER), can be efficient for information extraction tasks for information with clear patterns or structures [24]. However, when these patterns are not clear, or when these patterns are constantly changing, rule-based methods struggle with the flexibility required to handle these diverse texts [19]. On the other hand, machine learning methods for information extraction, particularly large language models based on transformer architectures such as Generative Pre-trained Transformers (GPTs) and Bidirectional Encoder Representations from Transformers (BERT), have high flexibility in understanding different contexts, but face challenges with scalability when extracting large volumes of information from large documents, as they require large amounts of labeled training data [19] or face issues processing large amounts of tokens [30]. GPTs, which do often not require training of the data for the task, may suffer from underfitting for specific tasks and require precise prompts to ensure efficient information extraction [13].

Papers that focus on information extraction (IE) often treat rule-based and machine learning methods in isolation, focusing on improving rule-based systems (e.g. [24]) or enhancing the contextual capabilities of machine learning models (e.g. [13]) for IE. Little research is dedicated to combining these methodologies to leverage their complementary strengths and reduce the influence of their limitations. This project aims to fill this gap by creating a hybrid system that combines traditional rule-based NER and RegEx with the advanced contextual understanding of the machine learning model ChatGPT.

This system is created by considering administrative decisions. These decisions are generally understood as an "administrative action addressed to one or more individualized public or private persons which is adopted unilaterally by a public authority to determine one or more concrete cases with legally binding effect"[18]. Examples of such administrative decisions include licensing, subsidizing, and sanctioning decisions [18]. Unlike other types of legal data, such as legislation and case law, administrative decisions have hardly been subjected to legal analytics. Administrative decisions are particularly suitable for this hybrid project, as the decisions are always subject to legalization with general requirements on the form and substance of these decisions. Every administrative decision should contain, for example, a date, a legal basis and a competent decision-making authority. At the same time, these decisions cover a rich variety of information types in which these general requirements are concretized[26]. This project will aim to set the first steps of quantitative legal analysis for administrative decisions by creating a hybrid system for efficient information extraction that is highly reproducible for different categories of administrative decisions.

To understand the efficiency of the hybrid system on administrative decisions, the research question that will be answered in this project is as follows:

RQ1*: How can a combination of traditional rule-based NER and RegEx with machine learning methods, more specifically ChatGPT, contribute to efficient information extraction from administrative decisions?*

To answer this research question, a deeper and nuanced understanding of what different types of information can be extracted from administrative decisions needs to be developed to achieve efficient information extraction from the hybrid model. Several sub-questions are therefore addressed:

- ***SQ1****: What types of information from administrative decisions can efficiently be extracted using rule-based methods only?*
- ***SQ2****: Regarding what types of information from administrative decisions can machine learning methods improve the information extraction from rule-based methods?*
- ***SQ3****: Regarding what types of information from administrative decisions can rule-based methods improve the information extraction from machine-learning methods?*

The results show that a combination of methods can be efficient for information extraction. Rule-based methods are efficient when patterns or structures are clear and do not require context-aware techniques. In addition, rule-based methods can be efficient in identifying key sentences to shorten the amount of text given to the machine-learning method, increasing its performance and achieving efficient information extraction from types of information that require extensive fine-tuning for an accurate extraction by rule-based methods. Information where identifiable patterns are missing can be extracted using machine-learning methods from the given context or sentences from information types that do consist of identifiable patterns, extracted by rule-based methods. Certain limitations still apply, for example, patterns by rule-based methods need to be correctly identified to achieve effective information extraction, and the sentences or context given to the machine-learning method need to be sufficient and complete, which in some cases may be hard to achieve with the chosen rule-based methods.

This paper is organized as follows. The introduction gives an introduction to the problem and introduces the research questions central to the paper. The related work section gives a brief overview of prior research in different tasks that apply different rule-based and machine-learning methods to legal information, which is followed by an overview of prior research on the relative performance of different types of IE techniques. The method explains in detail how the research questions will be answered, identifying the different types of information in administrative decisions, explaining the data collection and selection approaches used, after which the approach for rule-based and machine learning methods for each different type of information is explained. The results section shows the results for each type of information and shapes an answer to each of the sub-questions based on the obtained results. The discussion section will highlight influences or problems that may have influenced the results, including the generalizability and a brief discussion of certain ethical issues regarding information extraction on legal data such as administrative decisions. This paper ends with a conclusion, which summarizes key highlights and answers the research question.

2 RELATED WORK

2.1 Legal information extraction

2.1.1 The need for legal information extraction. The field of legal information extraction and its techniques have gained increased attention due to the role it plays in facilitating access to legal knowledge and helping legal professionals in their tasks. As the volume of legal documents continues to grow and as more legal documents are being published publicly, efficient and effective extraction of relevant information is of utmost importance. Researchers in the legal field have explored various techniques and methodologies aimed at automating the extraction process and enhancing the usability of legal texts. [19]. Oard et al. (2010) [16] describe the complexity of information extraction techniques for legal data, highlighting challenges for information extraction due to the volume, variety, and complexity of legal data. Additionally, they describe frameworks for information extraction and its evaluation, highlighting the need for correct annotation sets[16].

As highlighted by e.g. Zadgaonkar and Agrawal (2021) [28], there is a need for information extraction from legal data and it can be used to achieve various goals, such as analyzation of legal data and decision-making purposes. Legal information extraction differs from other information extraction tasks from other domains, as legal data often consists of longer documents, complex internal structures, and jargon [28]. This section aims to identify what kind of rule-based and machine-learning methods are used for which kind of tasks and what type of information.

2.1.2 Rule-based methods. Rule-based methods are often applied in contexts when extracting, classifying, or annotating textual data. In earlier years, rule-based methods such as NER were mainly used to extract information from legal data. Major steps have been taken in the identification of legal references such as laws and citations by applying techniques like NER on legal data [24]. Methods like Technology Assisted Review (TAR), in which information extraction plays a central role, have also been applied to legal data, which applies techniques similar to reinforcement learning from human feedback (RLHF) to detect and identify patterns in legal data for data categorization. This process was proven to be more time-efficient than training a model on pre-annotated data [10, 23]. Furthermore, research has been dedicated to fine-graining NER in legal documents [14]. Additionally, legal documents have been classified and visually simplified by creating a semantic network by using for example NER and POS-tagging [7]. The need to identify the different patterns in a flexible, precise, and correct way is a limitation that is commonly discussed when working with rule-based methods [24]. In conclusion, rule-based methods for IE tasks are often used in the legal field on data that consists of clear patterns, such as laws and references, and are further applied for classification and annotation tasks.

2.1.3 Machine learning methods. In more recent years, machine learning methods like LLMs have gained attention in the legal field because of their effectiveness for information extraction tasks. For example, GPTs like ChatGPT have been used in legal research for identification of legal factors in legal opinions [9], the summarization of legal contacts[30] and rhetorical role prediction in legal cases [1]. Other machine learning techniques have also been researched, for example, a variation of BERT was used on legal text for summarization [2], and similarly, summarization of judgment decisions was achieved by using nearest neighbor search [20]. However, studies like Sansone and Sperlí (2022) [19] show that little research is dedicated to the information extraction from administrative decisions as opposed to other legal documents, such as laws or court judgments. In conclusion, machine-learning methods are often applied when rule-based methods cannot be applied due to the absence of clear patterns, or for more complex tasks such as summarization.

2.2 Rule-based and machine learning techniques in general

2.2.1 Rule-based methods. For tasks outside of the legal domain, techniques like NER, in combination with for example BERT and Relation Detection (RE), are widely used and have been proven to be efficient methods for these tasks. For example, Chandramouli et al. (2021) applied a combination of NER with BERT on unlabeled transcripted audio data which leads to near-human accuracy for classification tasks [5]. Bui et al. (2016) successfully applied NER to extract information from different PDF documents, such as title and body text [4]. Additionally, this paper uses an evaluation approach that evolves around a so-called 'gold standard', where the accuracy is measured based on pre-labeled data, which is a common approach for machine extraction evaluation [4].

Other tasks in which rule-based methods are applied are described by scholars like Haak (2020), where a similar method is applied, and which concludes that BERT-based models can contribute to the effectiveness of NER or Semantic Role Labeling (SRL) [11]. This semantic role labeling is also applied in papers like Perera et al. (2020), where NER is used in combination with relation detection [17]. The paper discusses the potential of relation detection by using it to categorize topics of academic papers. Siciliani et al. (2023) discuss different types of techniques, more specifically different deep learning techniques for information extraction, such as XGBoost, which was proven to be effective for the extraction and classification of relations from tenders of the public sector [21].

2.2.2 Machine-learning methods. Recent studies have shown that generative large language models (LLM), such as ChatGPT, are on the rise for information extraction tasks and can lead to promising results with minimal resources [13]. The advantage of using these LLMs is that zero-shot learning can be applied to information retrieval. Studies show that near-human results can be achieved in this way, and may in some cases even lead to better accuracy's than certain full-shot models [25]. The main challenge for using ChatGPT lies in creating an effective prompt that leads to the best results. Studies have shown that ChatGPT is effective in extracting attributes and their relations for for example products [3].

Some researchers have pointed out limitations of using ChatGPT or other generative LLMs for information extraction tasks. Zhang et al. (2023) for example state that ChatGPT is effective for the extraction of relevant information, but is limited in retrieving more specific information [29]. In other words, the study shows how ChatGPT for the task of the study resulted in high recall but low precision. Additionally, studies like Tong and Chengzhi (2023) show how ChatGPT may lead to lower performances for different kinds of languages, in the case of Tong and Chengzhi (2023) for Chinese texts [22]. It is unclear how effective information extraction with ChatGPT is on Dutch texts.

Additionally, studies like Zin et al. (2023) indicate how ChatGPT is less effective when providing large amounts of text, and achieved accurate results when splitting the text in different prompts for the summarizing of legal contracts [30]. This is mainly due to the token limitations from ChatGPT and similar models, which makes it unable to process large amounts of text at the same time. Additionally, processing large amounts of text and thus making the model process a high amount of tokens may increase the price of prompts by including non-relevant information.

3 METHODOLOGY

This section describes the steps that have been taken to answer the research question and its sub-questions. This is done by considering Dutch administrative decisions. These are commonly not publicly available. In The Netherlands, steps are taken to achieve a more transparent government. In 2022, a new law on public disclosure of government documents has entered into force under Article 3.3 of this Dutch Open Government Act (Woo)[8]. This law will make it mandatory for governments to publicly disclose these decisions proactively[26]. Scholars in this field discuss how this kind of disclosure could enable individuals to compare their case with others [15]. Information extraction techniques can make this process easier to achieve, by allowing for a quantitative comparison between decisions[26].

Because of the considerable length and heterogeneous characteristics of (certain) administrative decisions[26, 28], the limitation of ChatGPT being able to efficiently process large documents[29, 30], and the increasing availability of Dutch administrative decisions, this project focuses on the combination of rule-based methods and ChatGPT as machine-learning method for Dutch administrative decisions. More specifically, SpaCy's[1] NER and POS techniques in combination with RegEx techniques, which were shown to be used in legal data for for example the extraction of references [24], will be applied to identify key sentences, which reduces the amount of input tokens necessary to obtain results from ChatGPT. A combination of these two will thus be created to enhance each method's strengths and limit their weaknesses.

3.1 Data collection and selection

To answer the research question, publicly disclosed administrative decisions for two government bodies have been used, namely the Kansspelautoriteit (Dutch gambling authority, KSA)[2] and the Autoriteit Financiële Markten (Dutch financial markets authority, AFM)[3]. This data is publicly available on the websites of the individual government bodies and on aggregated websites like Woogle[4]. The used data from Woogle is unstructured and unlabeled and contains different (unlabeled) types of decisions, such as licensing decisions and sanctioning decisions. The obtained data consists of various types of administrative decisions. Based on Article 3.3a of the Dutch Open Government Act[8], several information types are identified that are to be expected in every type of administrative decision, such as the data of the decision and the legal basis for this decision. This resulted in the different types of information as described in table 1. However, apart from these general characteristics applicable to all administrative decisions, each type of administrative decision also contains other types of information that is distinct for that particular type of administrative decisions. Therefore, this project focuses in particular on enforcement decisions, as these contain similar types of information. Enforcement decisions are administrative decisions in response to certain misconducts (i.e. violation of a legal provision). Administrative fines and administrative penalties are selected as enforcement decisions for this project. Administrative fines discuss an unconditional obligation to pay a sum of money, whereas administrative fines discuss a conditional obligation to pay a sum of money if the violation is not terminated within a certain period. Based on the General Administrative Law Act[8], other information types that are legally present in enforcement decisions have been identified, as seen in table 1. An example of a shortened administrative fine is displayed in figure 1, where the information to be extracted as described in table 1 is annotated based on the annotation protocol in appendix D.

[1] SpaCy's pre-trained pipeline 'nl_core_news_lg'.
[2] kansspelautoriteit.nl/
[3] afm.nl/
[4] woogle.wooverheid.nl; downloaded in April 2024.

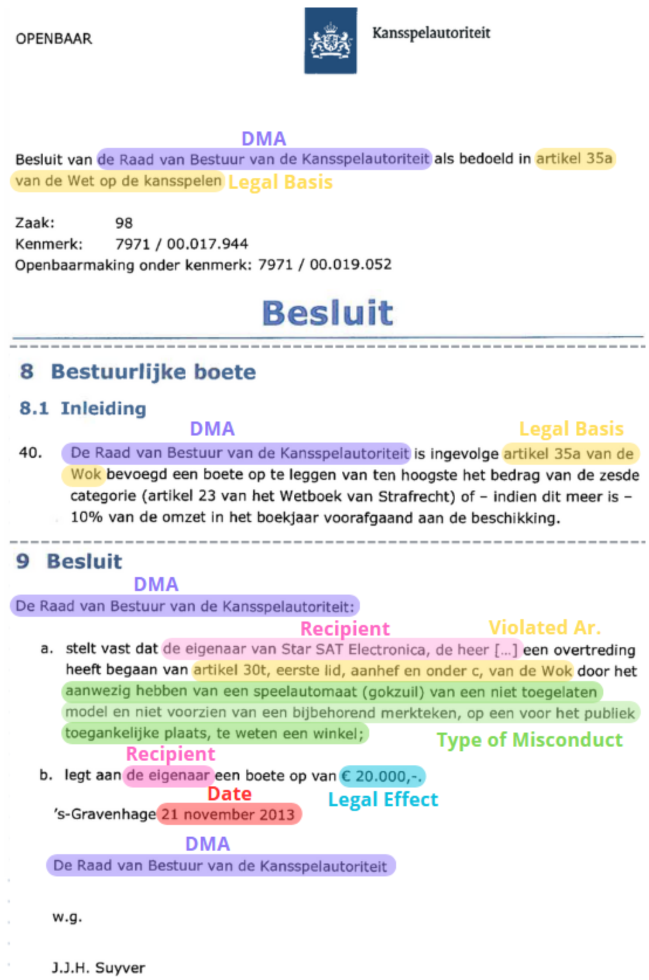
	Type	Description
All	Date	Date of decision.
	Legal Effect	Given sanction; money or consequence.
	Legal Basis	Legal provision; authority for DMA for making decision.
	Recipient	Legal entity; Person, organization.
	DMA	Decision Making Authority, Governing body making the decision.
fine, penalty	Violated Ar.	List of legal provision(s) that are violated.
	Type of Misconduct	Sentence containing misconduct.

Table 1: Information that should be present in all enforcement decisions[8].

The obtained data includes categories of administrative decisions that are irrelevant to this paper, such as licensing decisions. Data selection will be done to select only administrative fines and penalties within the set of administrative decisions available from both government bodies:

(1) *Keyword extraction.* A keyword extraction technique is applied to create a subsection based on present or absent keywords for both categories. These include 'decision', and for example 'fine' and 'penalty' for administrative fines and penalties respectively.
(2) *Remove irrelevant documents through keyword extraction.* Similarly to step 1, keywords that indicate an advice document are extracted, after which the document is shortened or completely removed. In some cases, documents include an advice in their appendix in a decision. By shortening the full document and focusing solely on the decision, future steps for information extraction may be improved.
(3) *Extraction of Legal Effect to remove and classify documents.* Lastly, the technique described in section 3.2.3 is applied to find the Legal Effect. If any obtained Legal Effect is associated with a word like 'fine' or 'penalty', these documents are classified with their corresponding category. If no legal effect or no matches with keywords are found, and there is no indication of the decisions resulting in a 'no fine' or 'warning' result, the document is not selected.

This results in a selection of administrative fines (267) and administrative penalties (171) (appendix B). However, this selection contains noise, as the data of Woogle also includes administrative decisions related to these enforcement decisions, such as disclosure decisions and possible decisions on appeal (in response to objection to an enforcement decision).


OPENBAAR

Kansspelautoriteit

Besluit van de Raad van Bestuur van de Kansspelautoriteit als bedoeld in artikel 35a van de Wet op de kansspelen

Zaak: 98
Kenmerk: 7971 / 00.017.944
Openbaarmaking onder kenmerk: 7971 / 00.019.052

Besluit

8 Bestuurlijke boete

8.1 Inleiding

40. De Raad van Bestuur van de Kansspelautoriteit is ingevolge artikel 35a van de Wok bevoegd een boete op te leggen van ten hoogste het bedrag van de zesde categorie (artikel 23 van het Wetboek van Strafrecht) of – indien dit meer is – 10% van de omzet in het boekjaar voorafgaand aan de beschikking.

9 Besluit

De Raad van Bestuur van de Kansspelautoriteit:

a. stelt vast dat de eigenaar van Star SAT Electronica, de heer [...] een overtreding heeft begaan van artikel 30t, eerste lid, aanhef en onder c, van de Wok door het aanwezig hebben van een speelautomaat (gokzuil) van een niet toegelaten model en niet voorzien van een bijbehorend merkteken, op een voor het publiek toegankelijke plaats, te weten een winkel;

b. legt aan de eigenaar een boete op van € 20.000,-.

's-Gravenhage 21 november 2013

De Raad van Bestuur van de Kansspelautoriteit

w.g.

J.J.H. Suyver

Figure 1: Example of a shortened administrative fine with the information as described in table 1 annotated based on the annotation protocol in appendix D.

To remove disclosure decisions, legal provisions are extracted from the document (see section 3.2.2). From this list of legal provisions, keyword matching is applied to check the presence of legal provisions that indicate a disclosure decision, such as 'article 3.1 from Woo'. If a match is found, the document is removed from the selection. Regarding appeal decisions, government bodies are required to include an option for the recipient to object to the decision. The time frame to send this objection is legally required to be six weeks and the type of objection differs from administrative fines/penalties and objection to appeal decisions. Sentences that contain the phrase 'six weeks' were extracted from the document, after which the sentence is checked to have the phrase 'notice of objection' present. If no sentence included this word, the document was dropped. This resulted in a selection of administrative fines and penalties, which is visualized in appendix B.

The selected data included 299 different documents with various structures from the two government bodies. This data has been analyzed, which is shown in figure 2 and also displayed in appendix

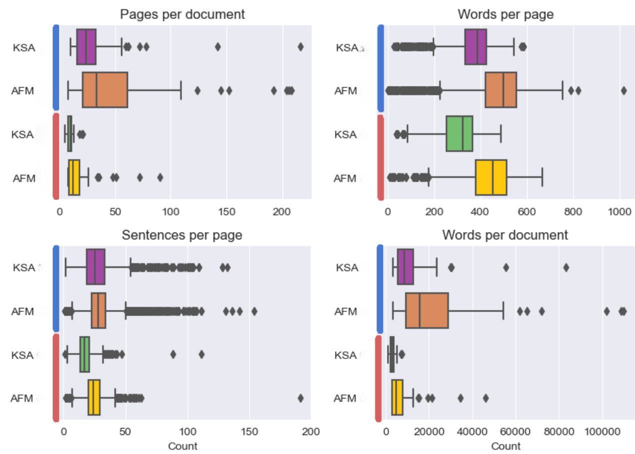


Figure 2: Document and page analysis for the two government bodies for administrative fine and penalty decisions. Blue and red indicate fines and penalties respectively.

	How	Result
Date	3.2.1. Date patterns, identify (1) recurring dates per page, or (2) dates in connection with a Dutch city or 'Date'.	**Date**
Violated Ar. and **Legal Basis**	3.2.2. Article patterns, POS-tagging to obtain a full article and apply keyword matching in the context of sentence to extract candidate sentences.	Candidate sentences
Legal Effect	3.2.3. Money-patterns, identify associated noun(s) using POS-tagging and select sentences where keywords match associated nouns.	Candidate sentences

Table 2: Information for which rule-based methods are applied to extract information or candidate sentences.

C. The documents are relatively long, containing many pages and words, which aligns with the lengthy characteristic described in section 2. Since the data is relatively clean, only identified headers and footers are removed from the text from each page, by removing common prefixes. These are however still saved to extract for example Date, as explained in section 3.2.1, as qualitative analysis shows that this type of information is often present in the header or footer of a document.

3.2 Rule-based: extracting candidate sentences

As mentioned in section 1 and 2, rule-based methods excel in the extraction of information when patterns are clear and structures are similar. Various parts of the information to be extracted from the data consist of identifiable patterns or structures, which are explained in table 2. For Date, the date pattern is identified and extracted after a few checks. No machine learning methods were applied to identify the date of the decision. For Violated Article, Legal Basis, and Legal Effect, patterns are identified, after which context-aware techniques are applied to the matches to identify if the pattern contains candidate information. Afterward, its sentence and neighbor sentences are extracted, which is included in the prompt for the machine-learning model, ChatGPT, which will analyze the sentences and extract the desired information (see section 3.3). This subsection explains the techniques used for each type of information that can easily be extracted through rule-based methods and the type of information that includes identifiable patterns but may require extensive pattern recognition to correctly extract information without noise or hallucinations. An overview of this subsection can be seen in table 2 and figure 3.

3.2.1 ***Date***. The date of the decision is the only type of information that is solely found by using rule-based methods (SQ1). Date included patterns that were uniform across documents, decisions, and government bodies. The following rule-based approach is applied to find the date of the decision:

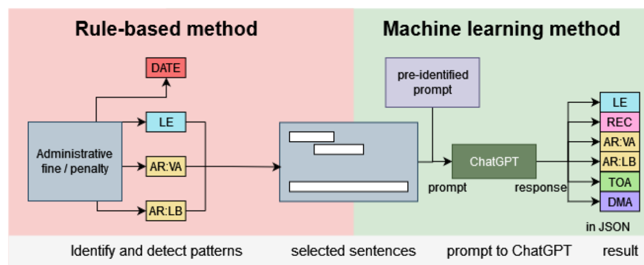


Figure 3: Pipeline of methodology.

- *Matching on date-patterns.* Firstly, the pre-trained NLP model was fine-tuned to detect dates and date patterns. This results in results with high recall but low precision, so additional DateTime-checks are applied to ensure the extracted match is a date.
- *Date presence.* In some decisions, the date of the decision is present on every page, for example, the header or footer. If an extracted date is thus present on all pages, it is likely the 'Date' is extracted, and the following steps are halted.
- *Keyword matching.* If no 'Date' was found in step 2, keywords were matched based on the context of the pattern (2 tokens before and after). If any token contains the word 'date', it is saved as a candidate date. Additionally, a database of Dutch cities[5] was used to check if any city is present in any context-token. This approach works for decisions that are written in letter format, where the date is often followed by a city, or vice versa. From all found candidate dates, the most recent date is chosen as 'Date'. If no candidate date was found, 'unknown' is returned for 'Date'.

[5] https://metatopos.dijkewijk.nl/

3.2.2 ***Violated Article and Legal Basis.*** These information types can be identified by rule-based methods, but require extensive pattern detection to correctly identify (SQ2). This is often due to the required context-aware characteristic of these data. To identify for example the violated legal provision, an analysis of the context needs to be applied.

The violated legal provision(s) (Violated Ar.) and the legal provision that explains the legal basis of the decision-making authority (Legal Basis) are recognizable as they contain a consistent pattern since they are a reference to a certain piece of legislation. An example of a legal provision is: 'article 30t, first paragraph, opening words and under c, of the Wok'. This pattern is consistent, as the word 'article' is always present and followed by a number (or a sequence of numbers) and/or letter, which is then linked to a law, in the example's case the Wok. In between the word 'article' and law, extra text can be added that specify what part of the article is being treated. Slight variations exist, such as naming multiple articles at the same time, often with the word 'juncto'. To extract the articles and their corresponding sentences, the following steps have been taken:

(1) *keyword matching*. The words 'article' and its multiplication are identified, after which its sentence is extracted. The sentence is cut short and starts at the keyword.
(2) *POS-tagging*. After extracting the (shortened) sentence, POS-tagging is applied to the sentence and is cut short at the first appearance of a token being identified as a verb, or until the end of the sentence is reached. This is based on a pre-trained NLP model on Dutch text. This approach is effective, as there are no verbs when describing articles, but are often linked to verbs (such as the verb 'is violated'). This shortened sentence acts as the possible article for Violated Ar. or Legal Basis.
(3) *Context-aware matching*. After identifying the article, the context of the article is taken into account. This is done by taking the sentence of which the article is part and 3 tokens from its neighbour sentences. This context is checked for words that indicate it is a violated article (Violated Ar.) or an article for the authority to make decisions (Legal Basis). For efficiency, verb-tokens are stemmed using SpaCy's tokenizer. For Violated Ar., these words include 'violate', and for Legal Basis, these include 'basis', and 'qualified'. If the context matches any of these, its sentence and its neighbor sentence are saved to be analyzed by the machine-learning method.

3.2.3 ***Legal Effect.*** Due to the choice to select administrative fines and penalties (as mentioned in section 3.1), the Legal Effect follows a similar pattern, as it always consists of a monetary amount, zero/nothing, or a warning. An example of a Legal Effect from an administrative fine is '€ 20.000,-', and for administrative penalty decisions '€ 1.000,- for each day until a maximum of € 10.000,-'. To identify the Legal Effect and extract its sentences for the machine learning method, the following steps have been taken:

(1) *Money-pattern matching*. By applying a combination of RegEx and NER, money patterns are being recognized and extracted. Since the pre-trained NLP model was deemed ineffective for correctly recognizing money instances through NER, RegEx has been added to the matcher to increase performance. For each match, the match and the sentence are extracted.

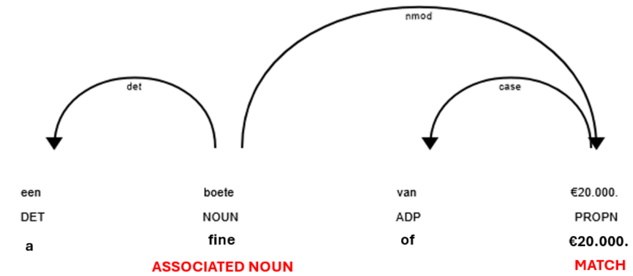


Figure 4: POS-tagging approach to identify context from money matches.

(2) *POS-tagging*. Similarly to section 3.2.2, POS-tagging is applied to the sentence, which shows the dependencies of the matched words. Through these dependencies, parent tokens from matched tokens are analysed, and the associated noun from the matched token is extracted, including extra information such as adjectives. See figure 4.
(3) *Keyword matching*. Based on the found associated nouns, the presence or absence of keywords is being analyzed. If the associated words include words like 'fine' or 'penalty', it is a candidate for Legal Effect. Additionally, if the associated words include words such as 'maximum' or 'basis', the found match likely indicates hypothetical or maximum penalties, and is thus removed. The sentences and their two neighbor sentences from the selection are saved for analysis by the machine-learning model.

In some decisions, the Legal Effect is a conclusion of not giving the recipient a penalty, or giving them a warning. To include these types of Legal Effects for analysis by the machine learning model, keyword matching has been applied, selecting the sentences (and their neighbors) that contain words such as 'warning' or 'no fine'.

3.3 Machine-learning method: analyzing sentences

This section deals with information types where rule-based methods can identify patterns, but require many resources to correctly extract (SQ2) and information types that lack identifiable patterns or structures (SQ3). Section 3.2 resulted in a selection of sentences that contain different types of information. These can be analyzed by the machine learning model in a zero-shot manner for more accurate extraction. When creating the annotation protocol (appendix D), certain information types were identified that lacked structure, such as Recipient and Type of Misconduct. However, these information types are often near other information types that do have identifiable patterns, such as Legal Effect and Violated Article. The information without clear patterns is thus included in the collected sentences and can be extracted efficiently by the machine learning method. Figure 3 gives a visual overview of the methodology, and what types of information are identified or extracted at what time.

This project uses ChatGPT as the machine-learning model, more specifically the model gpt-3.5-turbo-0125[6]. This language model

[6]https://platform.openai.com/docs/models/gpt-3-5-turbo

is based on the GPT-3.5 architecture, which utilizes transformer networks to analyze and generate text through extensive training on diverse datasets, including Dutch data. Its capabilities in natural language understanding make the model suitable for the information extraction task. This subsection explains the steps that are taken to obtain the extracted information from the GPT model from the candidate sentences, of which the prompts are also shown in table 3.

Firstly, all the sentences from the rule-based methods are collected together and ordered, which results in a list of candidate sentences from a particular administrative decision. If two sentences are next to each other in the document, they are added together so that they are coupled with each other. This allows for proximity sentences to be seen as part of each other, whereas a split indicates that text has been skipped. This improved the performance of the model. Afterward, the text generation method chat completions from the OpenAI API is used to interact with the model. A temperature of 0 is chosen, reducing randomness from the generation of the model. A pre-defined zero-shot prompt is given to this model which aims to extract the desired information for the two categories of decisions. For enforcement decisions, the prompt is as follows (translated into English):

> The given list of sentences originated from a single enforcement decision where the recipient(s) may have violated an article. Extract from the list of sentences for each recipient one-time the correct values for these keys in the following structure: 'fine': [{'Legal Effect': < table 3: Legal Effect Fine >, 'Recipient': < table 3: Recipient >, 'Violated Ar.': < table 3: Violated Ar. >}], 'Type of Misconduct': < table 3: Type of Misconduct >, 'DMA': < table 3: DMA >, 'Legal Basis': < table 3: Legal Basis >. Give your answer in JSON format. Sentences: [list_of_sentences]

The prompt differs slightly for the administrative penalty decisions. For example, the word 'penalty' is used instead of fine. This prompt is as follows:

> The given list of sentences originated from a single administrative penalty decision where the recipient(s) may have violated an article. Extract from the list of sentences for each recipient one-time the correct values for these keys in the following structure: 'penalty': [{'Legal Effect': < table 3: Legal Effect Penalty >, 'Recipient': < table 3: Recipient >, 'Violated Ar.': < table 3: Violated Ar. >}], 'Type of Misconduct': < table 3: Type of Misconduct >, 'DMA': < table 3: DMA >, 'Legal Basis': < table 3: Legal Basis >. Give your answer in JSON format. Sentences: [list_of_sentences]

As seen in both prompts, a value called 'fine' or 'penalty' exists, which shows a list of the values for Legal Effect, Recipient, and Violated Ar. This is to allow the model to extract multiple fines or penalties if multiple recipients are given one. Prior qualitative analysis highlighted this. The result of this prompt is a JSON file containing the extracted information for each type of information. This file is converted to a CSV file for evaluation. Additionally, Date is added to this CSV file, which was previously obtained as explained in section 3.2.1.

	Part of prompt
Legal Effect **Fine**	<amount (number) of the obtained fine. If it is decided to not give a fine, explain whether this is a 'warning' or '0'. No maximal, hypothetical, basis or fines from the past.>
Legal Effect **Penalty**	«amount (number) of the obtained penalty> per <unit> untill <maximum (number)>. If it is decided to not give an amount as a penalty, explain whether this is a 'warning' or '0'. No maximal, hypothetical, basis or penalties from the past.>
Recipient	<Legal entity that obtains the fine/warning, as complete as possible>
Violated Ar.	<[Which articles are being discussed if they are violated. Give each article in the following structure: article + number + possibly exordium + law]>
Type of Misconduct	<What has happened that the law/article is violated>
DMA	<Decision making authority; the governing body that is authorised to impose the fine>
Legal Basis	<on the basis on which article the DMA is authorized to take the decision. Give the article in the following structure: article + number + possibly exordium + law>

Table 3: Part of the prompt for each information type that has been used for the information extraction task.

3.4 Evaluation

After obtaining the extracted information from the GPT, the results are evaluated on precision, recall, and F1-score. This is a common method used in the field of information extraction [16]. The metrics provide a comprehensive understanding of the model's performance, where each metric focuses on different aspects of its accuracy and effectiveness. To evaluate the model a golden standard method is applied, for which a subset is hand-annotated based on an annotation protocol developed under the supervision of a legal domain expert (appendix D). The set consists of 10 documents for each combination of type and government body (=40 total).

To further ensure the reliability and validity of the evaluation, a percentage agreement score is calculated to measure the inter-rater reliability between two annotators, which is displayed in table 4. Percentage agreement was chosen as it gives a clear interpretation of the robustness of the annotation protocol and how reproducible

	Date	Legal Effect	Violated Ar.	Legal Basis	Recipient	DMA	Type of Misconduct
Percentage Agreement	100%	91.7%	100%	76.9%	83.3%	100%	90.9%

Table 4: Percentage agreement for each information type to assess inter-rater reliability for the golden annotated set based on appendix D.

the hand-annotated set is. Advanced techniques (such as Cohen's Kappa) that take to account chance agreement may not be necessary, as the extractions are based on a deeper understanding of the content and agreeing by chance is thus minimal. The percentage agreement was calculated by dividing the amount of agreements with the sum of agreements and disagreements. For Type of Misconduct, agreements were identified by hand on contextual agreements, allowing for paraphrased sentences to be seen as agreements.

A second masters student has independently annotated a subset of the golden set (10 documents) based on the annotation protocol to calculate the percentage agreement. Table 4 shows high agreement scores for all information types. The disagreements observed were due to differing interpretations by the annotators. One annotator identified a single legal basis, while the other identified two. Additionally, one annotator viewed some recipients as a single entity, whereas the other saw multiple recipients. These differences indicate improvements for the annotation model for future research for specific types of information. However, the overall strong agreements provide a greater confidence in the evaluation scores from the model. The machine-generated extraction is evaluated based on the golden set in the following way:

3.4.1 Precision. The precision shows the ratio of correctly extracted information from all the predicted extracted information. In other words, it is 'the proportion of retrieved documents which were relevant' [16]. For Legal Effect, Violated Ar. and Legal Basis, macro averaged precision is calculated by dividing the amount of correctly extracted information by the amount of extracted information.

For Recipient, Type of Misconduct, and DMA, Bilingual Evaluation Understudy (BLEU) scores have been calculated. It is widely used in natural language processing (NLP) for assessing the precision of machine translation systems [27]. Due to BLEU's ability to measure the quality of text generated by a model, its relevance extends to other NLP tasks, including information extraction, despite being designed for evaluating machine translation. BLEU evaluates the quality of the text by comparing the n-grams (sequences of n words) from the predicted extracted text with the ground truth extraction. It counts how many n-grams from the candidate text appear in the reference texts. A n-gram of 1 is used, as this allows for a verification that all important terms are included. This approach is applied since the text from these types of information is unstructured and allows for multiple correct notations.

3.4.2 Recall. The recall shows the ratio of correctly extracted information from all the golden standard extracted information, or the 'proportion of the extant relevant documents that were retrieved by the system' [16]. For Legal Effect, Violated Ar. and Legal Basis, macro averaged recall is calculated by dividing the amount of correctly extracted information by the amount of to be extracted information from the golden standard.

Similarly to section 3.4.1, due to the allowed different language use from the model, Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1) is being used to calculate the recall for Recipient, Type of Misconduct, and DMA. Unlike BLEU, ROUGE is designed to assess how well generated text captures the important content of a reference [27]. This approach is applied to calculate the recall to indicate if an extracted sentence is similar to the ground truth.

4 RESULTS

To answer the research questions, the approach from section 3 was applied to the documents from the two government bodies and the two types of enforcement decisions, which led to 175 documents for administrative fines (KSA: 65, AFM: 110) and 124 documents for administrative penalties (KSA: 55, AFM, 69). The documents are evaluated using the method described in section 3.4, based on a pre-annotated set of 40 documents (10 for each combination of government body and category). An example of machine-extracted information is shown in appendix A. The precision, recall, and F1-score are shown in table 5. This section discusses the results in light of the sub-questions as defined in section 1.

4.1 SQ1: Homogeneous patterns and context

4.1.1 Date. As explained in section 3, Date is the only information type that is extracted using rule-based methods only. As shown in table 5, Date shows high precision and recall scores. This indicates that the found patterns for Date accurately capture the date of the decision and are uniform over the decisions and government bodies. In conclusion, if patterns are identifiable, and uniform and do not require many context-aware techniques, rule-based methods are efficient for the extraction of these types of information.

4.2 SQ2: Homogeneous patterns, heterogeneous context

4.2.1 Legal Effect. As shown in table 5, Legal Effect is effectively extracted for administrative fines, showing a recall of 1 and a precision of 0.95 and 1 for the two government bodies. This score is significantly higher compared to administrative penalties, showing evaluation scores between 0.6 and 0.8 for Legal Effect. This indicates that rule-based methods are effective in identifying administrative fines and their Legal Effect while being less effective for administrative penalties.

4.2.2 Violated Article. For administrative fines, the results show a precision of 1 and a recall of on average 0.95, indicating effective information extraction (table 5). For administrative penalties, this score drops significantly, resulting in a score around 0.7 with a slightly higher recall. This may be explained because, for administrative penalties, the violation of an article may not be as clear as for

			Government Body					
			KSA			AFM		
Category	SQ	Type	Precision	Recall	F1-score	Precision	Recall	F1-score
Administrative Fine	1	Date	1.000	1.000	1.000	1.000	1.000	1.000
	2	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
	3	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
		Type of Misconduct	0.811	0.818	0.814	1.000	1.000	1.000
Administrative Penalty	1	Date	1.000	1.000	1.000	0.900	0.900	0.900
	2	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
	3	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
		Type of Misconduct	0.954	0.958	0.956	0.659	0.724	0.690

Table 5: Precision, Recall, and F1-scores for the two government bodies Kansspelautoriteit (KSA) and Autoriteit Financiële Markten (AFM) for the seven information types as defined in section 3.

administrative penalties, as the recipient has time to change its behavior. The language used to indicate the violated article may thus differ and may have not been captured by the rule-based method as described in section 3.2.2.

4.2.3 Legal Basis. The results show a precision and recall score of 1 for KSA (table 5). However, the scores for AFM are significantly lower, showing scores of only 0.2 and 0.3. This significant decrease could be explained by great language or structural differences in the documents between the two government bodies. The rule-based methods correctly identify candidate Legal Basis for KSA decisions, but this is ineffective for decisions from AFM.

Based on these results, a conclusion can be made that when patterns are identifiable but require many resources to optimize, machine-learning methods can help by applying context-aware techniques for the information extraction technique while not requiring many resources. However, rule-based limitations still apply, as patterns need to be sufficiently flexible to identify the information in different contexts, languages, and text structures.

4.3 SQ3: Heterogeneous patterns and context

4.3.1 Recipient. The evaluation scores are consistent across the different government bodies and categories, showing a ROUGE recall score of 0.8 or higher as seen in table 5. This indicates that the sentences generated from the rule-based methods on information types for SQ2 often include the correct recipient. Qualitative analysis shows that in certain cases a single recipient is seen as multiple. An explanation for this could be that the recipient is often referred to differently, by for example using abbreviations. This can explain the lower precision scores for Recipient, as the given Recipient may in some cases been extracted incomplete, or seen as multiple whereas it was one.

4.3.2 Decision Making Authority and Type of Misconduct. The evaluation scores of these scores are found using BLEU and ROUGE (see section 3.4). The precision and recall scores as seen in table 5 indicate that the extraction is efficient, showing scores of around 0.85. The recall score is often higher than the precision score, indicating that the model often extracts more incorrect information, but generally extracts what needs to be extracted. The given sentences do not always contain enough context for the model to correctly extract the information.

Based on these results, a conclusion can be made that rule-based methods can help machine learning methods by reducing the amount of text given to the model, while still being able to efficiently extract information from the shortened text. The machine learning method allows for information extraction from information types with heterogeneous patterns or structures. Rule-based limitations apply, as the extracted sentences should contain sufficient information for the machine learning method to correctly shape the context, which is seen for the Recipient.

5 DISCUSSION

5.1 Generalizability

This information extraction task was performed on two types of enforcement decisions, administrative fines and penalties. These two were chosen as they contain similar types of information (section 3.1). Other types of enforcement decisions, such as administrative coercions where recipients are required to change their behavior without the obligation to pay an amount of money, consist of different types of information, for example, Legal Effect does not contain money patterns, and thus requires greatly different pattern recognition techniques for Legal Effect.

The structure of enforcement decisions is however similar, as they identify a certain misconduct from a (legal) person, which makes this project generalizable over other types of enforcement decisions, although requiring different patterns for Legal Effect-extraction for certain types. Other types of administrative decisions, such as permits or financial grants, are greatly different in structure, as they do not take into account misconducts. Not only does an application on these decisions require a change in Legal Effect-detection, Violated Article needs to be adjusted as well, as the recipient of the decision has not violated a legal provision. Other types of information or legal provisions should be identified, such as the legal basis for the recipient to obtain the permit, to apply this approach to other types of administrative decisions. Future research is needed to identify whether these newly identified information types and their context include the types of information that are extracted solely by the machine learning model, such as recipient and type of activity, to understand the generalizability of this approach on a specific type of administrative decision.

Besides the information types extracted in this project, other types of information could also be found using a similar approach based on the specific characteristics of certain types of administrative decisions (as opposed to the general characteristics identified in the OGA) [8]. Similar approaches could be applied to different information types, but they may require different patterns and the used rule-based methods may need to be adjusted slightly. A thorough analysis of the information and its context is needed to identify if the information type requires the hybrid system as seen for e.g. Violated Article, or if rule-based methods are sufficient in extracting the information for example seen for Date.

5.2 Rule-based methods

The rule-based methods used in this project may not be the most efficient way to capture patterns in the data. Other scholars in this field have for example developed methods to effectively capture laws and/or articles (LinkeXtractor, [24]). Additionally, approaches for automatic pattern recognition for information extraction have been developed in for example Technology Assisted Review (TAR) for document classification [10]. However, these methods were not chosen, as they require either many resources to work effectively [23] or could not be run due to API and accessibility issues. Using these methods may however increase the performance of the information extraction techniques, as it applies rule-based techniques that have been proven to be very effective and thus may extract the correct sentences more accurately. They may also increase generalizability, as these methods are more flexible and trained over different types of legal data or administrative decisions. However, to answer the research question, these methods are not required to obtain a sufficient understanding of the hybrid model. Future research can identify the influence of more advanced rule-based methods on information extraction in the hybrid system. Additionally, future research can apply technologies like TAR for the categorization process of administrative decisions as described in section 3.1, as these are expected to effectively categorize the entire dataset of administrative decisions, requiring no manual pattern recognition [10]. This may be especially useful when more administrative decisions are published under the Woo [8] in the Netherlands[26], or when similar movements in public disclosure of administrative decisions are happening in different countries.

5.3 Ethical issues regarding information extraction on legal data

When applying techniques as described above that automate pattern recognition or when automating the information extraction process by including machine learning methods, scholars like Hildebrandt (2012) [12] describe how this automatic extraction or categorization of legal data could influence decision-making by governing bodies or judges, as they could be influenced by the extracted data. Machine-learning methods or automatic pattern recognition like TAR [23] could identify patterns in the data that can not be identified by humans. This pattern recognition could influence the decision-making process of the judge, as the machine would show the importance of the focus on a pattern found by the machine itself. This could make the decision more machine-driven, but the desirability of which is unknown. [12].

Moreover, collected sentences from rule-based methods are sent to ChatGPT, the machine-learning method. Due to the deep learning nature of the GPT, ChatGPT likely trains its model on user input. This may cause privacy issues, as the model extracts for example names of recipients and what type of misconduct they have done. This sensitive information could be trained on and could be linked to other data, making it easier to identify the recipient and link this to the misconduct and decisions. By using local types of LLMs as the machine learning method, such as Meta's LLama[7], which has been applied for similar tasks[6], these problems can be countered and may be more suitable for documents that are not publicly disclosed.

6 CONCLUSION

This project applied a combination of rule-based methods with machine learning methods for information extraction on Dutch administrative decisions. To achieve this efficiently, the nature of the different types of information in these decisions have been analyzed. Rule-based methods serve to identify or extract types of information where patterns or structures are homogeneous. Machine learning methods can be used for the extraction of information types that require more context-aware techniques to accurately extract, or information types that contain heterogeneous patterns. Rule-based methods serve as a tool to reduce the amount of text that needs to be processed by the machine learning method. However, rule-based limitations still apply, as the machine learning method is dependent on the recall performance of the rule-based methods when extracting information.

In conclusion, a combination of methods can make the information extraction task more efficient, as it enhances the strengths of each method, reducing the amount of text needed for information extraction, while reducing their weaknesses, allowing for context-aware extraction without the use of many resources and allowing efficient information extraction from large bodies of text.

[7] https://llama.meta.com/

REFERENCES

[1] Belfathi, A., Hernandez, N., & Monceaux, L. (2023). Harnessing GPT-3.5-turbo for rhetorical role prediction in legal cases. *arXiv preprint arXiv:2310.17413.*

[2] Benedetto, I., Cagliero, L., Tarasconi, F., Giacalone, G., & Bernini, C. (2023). Benchmarking abstractive models for italian legal news summarization. In *Legal Knowledge and Information Systems* (pp. 311–316). IOS Press.

[3] Brinkmann, A., Shraga, R., Der, R. C., & Bizer, C. (2023). Product information extraction using ChatGPT. *arXiv preprint arXiv:2306.14921.*

[4] Bui, D. D. A., Del Fiol, G., & Jonnalagadda, S. (2016). PDF text classification to leverage information extraction from publication reports. *Journal of biomedical informatics*, 61, 141–148.

[5] Chandramouli, A., Shukla, S., Nair, N., Purohit, S., Pandey, S., & Dandu, M. M. K. (2021). Unsupervised paradigm for information extraction from transcripts using BERT. *arXiv preprint arXiv:2110.00949.*

[6] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1), 1418.

[7] Giri, R., Porwal, Y., Shukla, V., Chadha, P., & Kaushal, R. (2017). Approaches for information retrieval in legal documents. In *2017 Tenth International Conference on Contemporary Computing (IC3)* (pp. 1–6).: IEEE.

[8] Government of the Netherlands (No Date). Dutch open government act. Article 3.3a.

[9] Gray, M., Savelka, J., Oliver, W., & Ashley, K. (2023). Can GPT alleviate the burden of annotation? In *Legal Knowledge and Information Systems* (pp. 157–166). IOS Press.

[10] Grossman, M. R. & Cormack, G. (2016). Continuous active learning for TAR. *The Journal*, 4(3), 1–7.

[11] Haak, B. (2020). Information extraction from homicide-related dutch texts using BERT. *Jheronimus Academy of Data Science.*

[12] Hildebrandt, M. (2012). The meaning and the mining of legal texts. In *Understanding Digital Humanities* (pp. 145–160). Springer.

[13] Hu, D., Liu, B., Zhu, X., Lu, X., & Wu, N. (2024). Zero-shot information extraction from radiological reports using ChatGPT. *International Journal of Medical Informatics*, 183, 105321.

[14] Leitner, E., Rehm, G., & Moreno-Schneider, J. (2019). Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems* (pp. 272–287).: Springer.

[15] Noveck, B. S. (2017). Rights-based and tech-driven: Open data, freedom of information, and the future of government transparency. *Yale Hum. Rts. & Dev. LJ*, 19, 1.

[16] Oard, D. W., Baron, J. R., Hedin, B., Lewis, D. D., & Tomlinson, S. (2010). Evaluation of information retrieval for E-discovery. *Artificial Intelligence and Law*, 18, 347–386.

[17] Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8, 673.

[18] Research Network on European Administrative Law (2014). ReNEUAL model rules on EU administrative procedure. https://www.reneual.eu/images/Home/ReNEUAL-Model_Rules-Compilation_BooksI_VI_2014-09-03.pdf.

[19] Sansone, C. & Sperlí, G. (2022). Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106, 101967.

[20] Shimbo, A., Sugawara, Y., Yamada, H., & Tokunaga, T. (2023). Nearest neighbor search for summarization of japanese judgment documents. In *Legal Knowledge and Information Systems* (pp. 335–340). IOS Press.

[21] Siciliani, L., Ghizzota, E., Basile, P., & Lops, P. (2023). OIE4PA: Open information extraction for the public administration. *Journal of Intelligent Information Systems*, (pp. 1–22).

[22] Tong, B. & Chengzhi, Z. (2023). Extracting chinese information with ChatGPT: An empirical study by three typical tasks. *Data Analysis and Knowledge Discovery*, 7(9), 1–11.

[23] Tredennick, J. (2015). TAR for smart people. *Catalyst.*

[24] van Opijnen, M., Verwer, N., & Meijer, J. (2015). 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).*

[25] Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al. (2023). Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205.*

[26] Wolswinkel, C. J. (2024). Actieve openbaarmaking van beschikkingen. In *Nederlands Juristenblad*, volume 24 (pp. 1851–1857).

[27] Yang, A., Liu, K., Liu, J., Lyu, Y., & Li, S. (2018). Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. *arXiv preprint arXiv:1806.03578.*

[28] Zadgaonkar, A. V. & Agrawal, A. J. (2021). An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(6).

[29] Zhang, J., Chen, Y., Niu, N., & Liu, C. (2023). A preliminary evaluation of ChatGPT in requirements information retrieval. *arXiv preprint arXiv:2304.12562.*

[30] Zin, M. M., Nguyen, H. T., Satoh, K., Sugawara, S., & Nishino, F. (2023). Information extraction from lengthy legal contracts: Leveraging query-based summarization and GPT-3.5. In *Legal Knowledge and Information Systems* (pp. 177–186). IOS Press.

7 APPENDICES

Appendix A MACHINE GENERATED RESULTS

Legal Effect	Recipient	Legal Basis	Type of Misconduct	DMA	Legal Basis	Date
7500	'thee- en koffiehuis Kayseri'	['artikel 30t, eerste lid, aanhef en onder c, van de Wok']	Het aanwezig hebben van een speelautomaat van een niet toegelaten model en niet voorzien van een bijbehorend merkteken op een voor het publiek toegankelijke plaats	raad van bestuur van de Kansspelautoriteit	['artikel 35a van de Wok']	09/04/2015
500000	'N1 Interactive Limited te Malta'	['artikel 1, eerste lid, onder a, van de Wet op de kansspelen']	Via de website www.betchan.com zijn in elk geval in de periode 9 januari 2020 tot en met 9 september 2020 kansspelen online zonder vergunning aangeboden op – in elk geval mede – de Nederlandse markt.	raad van bestuur	['artikel 35a van de Wok']	30/03/2021
180000	'Come On Europe Limited (thans Co-Gaming Limited)'	['artikel 1, eerste lid, aanhef en onder a, van de Wok']	Het aanbieden van kansspelen zonder vergunning	Raad van bestuur van de Kansspelautoriteit	['artikel 35a van de Wok']	22/12/2014
7500	'Stichting en de heer [betrokkene]'	['artikel 30t, eerste lid, onder c, van de Wok']	Het aanwezig hebben van een speelautomaat, te weten een gokzuil, van een niet toegelaten model en niet voorzien van een bijbehorend merkteken, op een voor het publiek toegankelijke plaats	Raad van Bestuur van de Kansspelautoriteit	['artikel 35a van de Wok']	10/12/2014

Table 6: Example of machine extracted information for an administrative fine from governing body kansspelautoriteit (KSA) on 4 different documents.

Legal Effect	Recipient	Legal Basis	Type of Misconduct	DMA	Legal Basis	Date
2000 per dag tot 20000	'Zeker van Zanten'	['artikel 5:20 Awb']	Niet voldoen aan informatieverzoeken van de AFM	AFM	['artikel 5:20 Awb']	10/06/2021
2000000 per keer tot 2000000	'Friendly Finance B.V.'	['artikel 2:60, eerste lid, van de Wet op het financieel toezicht (Wft)']	Aanbieden van krediet zonder de vereiste vergunning	Autoriteit Financiële Markten (AFM)	['artikel 2:60, eerste lid, van de Wet op het financieel toezicht (Wft)']	12/07/2013
2000 per dag tot 20000	'Staten Assurantiën B.V.'	['artikel 2 Wet op het financieel toezicht']	Niet voldoen aan opgelegde last	Autoriteit Financiële Markten (AFM)	['artikel 1:79 Wet op het financieel toezicht']	14/03/2014
5000 per dag tot 50000	'N.V. Esperite N.V."	['artikel 5:33, eerste lid, onder a, sub I van de IVft']	Niet verstrekken van een schriftelijk overzicht van transacties in financiële instrumenten	Autoriteit Financiële Markten (AFM)	['artikel 1:1 Vft']	08/02/2019

Table 7: Example of machine extracted information for an administrative penalties from governing body Autoriteit Financiële Markten (AFM) on 4 different documents.

Appendix B DOCUMENT CLASSIFICATION

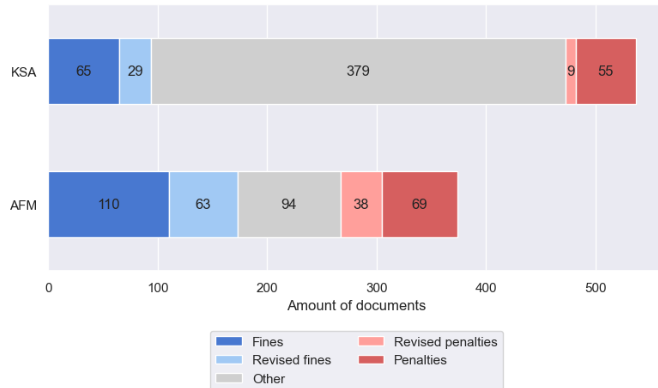


Figure 5: Document classification for administrative decisions for both governing bodies. Blue indicates administrative fine, and red administrative penalty decisions. A light color indicates an internal appeal decision.

Appendix C DOCUMENT ANALYSIS

	Pages per document	Words per page	Words per document	Sentences per Page
min	10	26	3396	2
mean	32.6	369.2	12059.0	27.8
median	24	385	8457	25
max	216	586	83112	132

(a) KSA fine (n=65)

	Pages per document	Words per page	Words per document	Sentences per Page
min	5	36	1051	1
mean	25.8	355.2	9177.8	24.5
median	11	372	3284	22
max	216	544	83112	111

(b) KSA penalty (n=55)

	Pages per document	Words per page	Words per document	Sentences per Page
min	8	1	3494	1
mean	48.7	478.6	23218.5	30.2
median	34	500	16286	28
max	208	1015	109688	154

(c) AFM fine (n=110)

	Pages per document	Words per page	Words per document	Sentences per Page
min	8	1	2877	1
mean	40.7	473.0	19206.1	30.8
median	18	489	7351	28
max	208	753	109688	191

(d) AFM penalty (n=69)

Figure 6: Analysation-scores for documents.

Appendix D ANNOTATION PROTOCOL

Note: The annotation guidelines are written in Dutch, since the administrative decisions are in Dutch.

Annotatie Protocol voor Informatie Extractie uit Beschikkingen

1. Doel

Dit annotatie protocol richt zich op het systematisch extraheren van belangrijke juridische informatie uit bepaalde soorten beschikkingen, namelijk uit boetebesluiten en dwangsombesluiten. Dit protocol wordt gebruikt om machine geextraheerde informatie te analyseren en evalueren.

2. Documenten

De documenten waar informatie uit gehaald moet worden zijn twee soorten sanctiebesluiten van twee verschillende bestuursorganen. De documenten zijn publiekelijk beschikbaar op woogle.wooverheid.nl. De bestuursorganen die voor dit project gebruikt worden zijn:

- **Autoriteit Financiële Markten (AFM).** Dit bestuursorgaan handhaaft en houdt toezicht op de integriteit en transparantie van de financiële markten in Nederland.
- **Kansspelautoriteit (KSA).** Dit bestuursorgaan ziet toe op de naleving van wet- en regelgeving binnen de kansspelsector in Nederland.

De volgende soort sanctiebesluiten kun je verwachten:

- **Boetebesluit**. Dit is een besluit waarbij een bestuursorgaan als bestuurlijke sanctie een onvoorwaardelijke verplichting tot betaling van een geldsom oplegt aan een persoon of entiteit vanwege een overtreding van wettelijke voorschriften. Er kan ook besloten worden om een waarschuwing of een boete van 0 euro op te leggen.
- **Last onder dwangsom.** Dit is een besluit waarbij een bestuursorgaan als bestuurlijke sanctie een maatregel oplegt waarbij de overtreder de last krijgt om een overtreding binnen een gestelde termijn te beëindigen, bij gebreke waarvan hij een geldsom is verschuldigd.

3. Richtlijnen voor annotatie

De volgende richtlijnen voor annotatie moeten worden gevolgd:

- De informatie die wordt geëxtraheerd is direct gekopieerd vanuit het document en in één stuk aan elkaar te vinden.
  - Stukken combineren uit verschillende stukken tekst kan dus niet.
  - Spelfouten of dergelijke zitten dus ook in de extractie, als deze in het document voorkomen
  - Bij een page break of dergelijke, moet de header, footer, paginanummer en dergelijke niet worden meegenomen.

- Voorbeeld van een **juiste** annotatie van Type Overtreding (en Overtreden Artikel):

- Voorbeeld van een **onjuiste** annotatie van Type Overtreding (en Overtreden Artikel):

- De geëxtraheerde informatie is opgehaald vanuit de juiste context.
  - Als de informatie op een ander stuk in het document vollediger is, maar niet in de context van de te annoteren informatie staat, kan deze niet worden gebruikt.
    - Voorbeeld: Als de wet van het overtreden artikel vollediger wordt weergegeven in een context waarbij niet duidelijk is dat het artikel wordt overtreden, kan deze niet worden gebruikt.
    - Voorbeeld: Het type overtreding wordt vollediger gemeld op een stuk waarbij niet duidelijk wordt vermeld dat deze handeling een overtreding is van de wet. De minder volledige versie in de juiste context wordt geëxtraheerd.
  - Voorbeeld van een **juiste** annotatie van overtreden artikel, vanwege de aanwezigheid van de (juiste) context (in dit geval: in stijd met, overtreding). Hoewel het artikel van de wet 'Wok' eerder in het document vollediger vermeld is (Wet op de Kansspelen), wordt deze niet geëxtraheerd omdat de volledigheid niet in de juiste context is.

- Elk uniek stuk informatie wordt eenmalig geextraheerd als stukken informatie meerdere malen in het document worden herhaald.
  - Voorbeeld: Ondanks dat de unieke ontvanger 'Clipboard Publications B.V.' meerdere malen in het document wordt vermeld, wordt de ontvanger slechts eenmalig geextraheerd.

  -

903

- Als er meerdere verschillende ontvangers, juridische effecten of artikelen worden benoemd, wordt elke unieke waarde geextraheerd.
    - Wel: verschillende waardes, andere ontvangers etc.
    - Niet: synoniemen, afkortingen etc.
- Het aantal boetes of dwangsommen dat wordt geextraheerd in een beschikking is altijd gelijk aan het aantal ontvangers.
    - Voorbeeld: [boete 1, boete 2] , [ontvanger 1, ontvanger 2].

4. Te Annoteren Informatie

De volgende informatie moet worden gelabeld en geëxtraheerd uit de beschikkingen, op basis van de richtlijnen gegeven in 3:

1. **Type sanctiebesluit**
    - **Beschrijving:** Het type sanctiebesluit dat in het document aan de orde is. Dit is of ‘Boetebesluit’, of ‘Dwangsom’.
    - **Annotatie**: Categoriseer het document als een ‘Boetebesluit’ of ‘Dwangsom’. Zie sectie 2 voor extra informatie.

2. **Datum (Date)**
    - **Beschrijving:** De datum waarop de beschikking is gegeven
    - **Annotatie:** Extraheer de datum waarop het besluit is genomen. Annoteer de datum in het formaat DD/MM/JJJJ.
        1. Voorbeeld**:** 25/01/2021
        2. Als de datum van het besluit ontbreekt, geef dit weer als ‘UNKNOWN’.

3. **Ontvanger (Recipient)**
    - **Beschrijving:** De persoon of entiteit die het sanctiebesluit ontvangt. Er kunnen meerdere ontvangers zijn.
    - **Annotatie:** Annoteer de naam van de ontvanger(s), zo volledig mogelijk.
        1. Als de ontvanger een persoon is, extract waar mogelijk ook de affiliatie of organisatie van deze persoon
            1. Voorbeeld: ‘Jan Smit, eigenaar van FC Volendam’.
        2. Als de ontvanger een B.V. of bedrijf is, geef deze zo volledig mogelijk weer.
            1. Voorbeeld: ‘Accountants Baat B.V.’
        3. Soms is de ontvanger geanonimiseerd. Extract dan de geanonimiseerde versie van de ontvanger
            1. Voorbeeld: ‘De heer [...]’
        4. Als er synoniemen voor dezelfde ontvanger wordt gebruikt, extraheer dan enkel de meest volledige versie die in de juiste context te vinden is.
        5. Als er meerdere ontvangers zijn, gebruik dan een apart veld voor elke ontvanger in de volgende vorm: [‘ontvanger 1’, ‘ontvanger 2’].

904

4. **Juridisch Effect (Legal Effect)**
    - **Beschrijving:** Het juridische effect (rechtsgevolg) van de beschikking, de juridische consequentie van het sanctiebesluit voor de ontvanger. Er kunnen meerdere juridische effecten zijn voor verschillende ontvangers.
    - **Annotatie:** Extract het juridisch effect van de beschikking. Het juridisch effect verschilt voor een boetebesluit en een sanctiebesluit:
        1. *Boetebesluit*. Extraheer het getal dat wordt opgelegd als boete. (NB: op basis van onderdeel 1 is al duidelijk dat sprake is van een boete). Als er een waarschuwing wordt gegeven, of als er wordt besloten om geen boete op te leggen, geef dit dan weer als ‘Waarschuwing’, of ‘0’.
            1. Voorbeeld: 10000
        2. *Dwangsom*. Extraheer het getal dat wordt gegeven als de hoogte van de dwangsom, per eenheid die wordt gegeven tot een maximum. Geef dit weer in de volgende vorm: ‘*Hoogte* per *Eenheid* tot *Maximum*’. Deze stukken hoeven niet aan elkaar in het document voor te komen.
            1. Voorbeeld: 2500 per overtreding tot 25000

            Hieronder staan de punten verder uitgewerkt:

            2. *Hoogte*: Een getal van een bedrag dat moet worden bepaald per eenheid
                1. Voorbeeld: 2500
            3. *Eenheid*: Een eenheid (zoals tijd, overtreding) waarbij de ontvanger de hoogte moet bepalen per strekking van de gegeven eenheid
                1. Voorbeeld: ‘per dag’, ‘per overtreding’
            4. *Maximum*: Het maximum aantal dat betaald moet worden als de overtreding niet wordt gestaakt.
                1. Voorbeeld: 25000

            Als een onderdeel mist, vul deze dan niet in.

            5. Voorbeeld: [‘2500 per overtreding’, ‘2500 tot 25000’]

        3. Als er meerdere juridische effecten zijn, geef dit dan weer in de volgende vorm: [‘juridisch effect 1’, ‘juridisch effect 2’]

5. **Overtreden Artikel (Violated Article)**
    - **Beschrijving:** Het artikel van de wet dat is overtreden. Er kunnen meerdere overtreden artikelen zijn.
    - **Annotatie:** Annoteer het artikel op de volgende manier: ‘artikel *Toevoeging* van *Wet*’.
        1. Voorbeeld van overtreden artikel: Artikel 33a van de Wet op de Kansspelen

    Hieronder worden de schuingedrukte termen verder uitgelicht:

2. *Toevoeging*. Het nummer van het artikel waarnaar gerefereerd wordt. Dit is vaak een getal, soms gevolgd door een nummer of met een speciaal karakter (zoals :). Neem aanheffen etc. mee in de annotatie.
    1. Voorbeeld: 2:60, eerste lid, aanhef en onder c
3. *Wet*. Dit is de wet waar het artikel zich in bevind. Vaak wordt dit weergegeven na het woord 'van'.
    1. Voorbeeld: Wet op het financieel toezicht

4. Als er meerdere overtreden artikelen zijn, geef dit weer als: ['overtreden artikel 1', 'overtreden artikel 2']
    1. Als er gebruikt wordt gemaakt van 'Juncto' (in combinatie met), extract deze dan als één artikel
        1.

6. **Besluitvormende Autoriteit (Decision Making Authority - DMA)**
    - **Beschrijving:** De autoriteit die de boete of dwangsom oplegt.
    - **Annotatie:** Annoteer de naam van de besluitvormende autoriteit of bestuursorgaan. Doe dit zo volledig mogelijk.
        1. Voorbeeld van besluitvormende autoriteit: raad van bestuur van de Kansspelautoriteit

7. **Juridische Basis (Legal Basis)**
    - **Beschrijving:** Het wetsartikel dat de besluitvormende autoriteit (DMA) de bevoegdheid geeft om het sanctiebesluit te nemen.
    - **Annotatie:** Annoteer het artikel op eenzelfde manier als Overtreden Artikel (Violated Article)
        1. Voorbeeld van juridische basis: artikel 35a van de Wet op de Kansspelen

8. **Type Overtreding (Type of Misconduct)**
    - **Beschrijving:** Beschrijving van de handeling(en) die geleid hebben tot de overtreding van het artikel.
    - **Annotatie:** Extraheer het stuk dat de feitelijke gedraging in detail beschrijft, wat er toe heeft geleid dat een wetsartikel is overtreden. Extraheer hiervoor de zin(nen) of gedeelte van de zin zo volledig mogelijk
        1. Voorbeeld van type overtredingen:
            1. Het aanbieden van gelegenheid tot gokken op sportwedstrijden op een niet toegelaten speelautomaat op een publiek toegankelijke plaats
            2. Het online aanbieden van kansspelen zonder vergunning
            3. Niet voldoen aan informatieverzoeken van de AFM

906