KLEROS

Enhancing Online Dispute Resolution through Natural Language Processing: **A Case Study of Kleros**

Alesia Zhuk

Universitat Pompeu Fabra
alesia.zhuk@ug.uchile.cl

Abstract

The growing significance of Online Dispute Resolution (ODR) lies in its capacity to provide efficient, accessible, and cost-effective solutions for resolving disputes outside traditional courtrooms. This study focuses on enhancing the functionality of Kleros, an innovative ODR platform, by integrating Natural Language Processing (NLP). Kleros faces challenges such as procedural inefficiencies, user comprehension difficulties, and the interpretation of complex legal terms, which hinder its broader adoption. By leveraging NLP, this paper proposes solutions to automate case analysis, simplify legal jargon, provide contextual explanations, and enhance the interpretation of user intent. These improvements aim to make the dispute resolution process more efficient and accessible for all parties involved. The findings highlight how NLP-driven enhancements can streamline ODR processes, improve juror experiences, and expand Kleros' applicability to a broader range of disputes.

Acknowledgement

The author would like to thank the Kleros Fellowship of Justice for providing the necessary support to conduct this research, and in particular Federico Ast and Robert Dean for generously sharing their insights.

1. Introduction

Alternative Dispute Resolution (ADR) has gained recognition as an alternative to the inefficiencies and high costs associated with traditional court systems (Carneiro et al. 2014, p. 212). Within ADR, Online Dispute Resolution (ODR) specifically refers to dispute resolution processes conducted entirely online, typically facilitated by private entities. While ADR is commonly associated with resolving disputes between private parties, it can also extend to state-related conflicts, such as investor-state disputes or disputes between states. In investor-state disputes, private entities (often corporations) seek to resolve conflicts with a state, typically concerning issues like expropriation, regulatory changes, or breaches of international investment agreements. These disputes are often settled through international arbitration, which is considered a form of ADR. Similarly, state-to-state disputes, such as those arising from territorial claims or treaty violations, may be addressed through ADR mechanisms such as diplomatic negotiation or international arbitration, enabling states to resolve conflicts without resorting to traditional court-based litigation or military confrontation. However, for the purposes of this paper, the primary focus will remain on ADR in the context of private-party disputes, where adjudication is carried out by private entities.

ODR, although a subset of ADR, is sometimes broadly used to describe any online dispute resolution method, including those employed by both private and state entities. It is crucial to differentiate between ODR and virtual hearings conducted by courts. Virtual hearings are part of the digital transformation of traditional court proceedings, yet they remain rooted in conventional judicial systems. In contrast, ODR operates outside this framework, typically characterised by its fully online nature and reliance on private-sector mechanisms for dispute resolution. While Alessa (2022, p. 320) refers to Lord Justice Briggs' proposal for an online civil court system within the Judiciary of England and Wales, such systems remain within the traditional judicial framework, adapting to online platforms rather than constituting a separate alternative system like ODR. The integration of digital tools into judicial processes does not equate to the creation of ODR systems, which are designed to offer an alternative to traditional courts.

Kleros is a prominent example of a modern ODR system, designed to resolve disputes in a decentralised manner using blockchain technology, particularly Ethereum, for jury selection and dispute allocation. Ethereum enables the use of smart contracts, so once evidence is reviewed and a majority decision is reached, the outcome is automatically executed. This system benefits the winning party and the jurors who support the majority decision, while penalising the losing party and those jurors whose decisions were in the minority, based on reputation or financial stakes if applicable. This decentralised, blockchain-based approach offers enhanced efficiency in terms of time, decision-making, and implementation compared to traditional dispute resolution methods. However, there are still areas that could be improved. The jury, composed of individuals who assess all the evidence presented by the parties, can face challenges in processing large volumes of information, potentially slowing down the decision-making process. Additionally, as jurors in Kleros are not expert judges but individuals from various backgrounds, this raises concerns about their ability to make fully informed decisions in more complex cases, which could impact the overall quality and accuracy of the resolution.

One potential area of improvement for ODR systems, such as Kleros, is the integration of more advanced technologies, particularly artificial intelligence (AI). Such advancements have the potential to not only refine the dispute resolution process from a technical perspective but also to improve overall outcomes, thereby increasing disputing parties' satisfaction (Carneiro et al. 2014, p. 212). While some authors suggest that ODR and AI have developed concurrently, with AI playing an increasing role in shaping ODR systems (Alessa 2022, p. 320), this perspective overlooks key differences in their development timelines. ODR systems began to emerge in the late 1990s, largely in response to the internet's rapid growth and the need for online mechanisms to resolve disputes efficiently. Early ODR systems were rudimentary, focusing on relatively simple processes like e-commerce dispute resolution. On the other hand, AI, while being a field of research for decades, did not see widespread practical applications until the 2010s, with its more extensive use in the 2020s, particularly in sectors like data analysis, machine learning (ML), and decision-making.

In the context of ODR, AI's influence has only become more pronounced recently. AI has the potential to significantly enhance ODR systems by improving efficiency, reducing costs, and providing more robust data analysis. However, its role at this stage is primarily about augmenting human decision-making rather than replacing it entirely. For example, AI could assist Kleros jurors by reviewing large volumes of evidence, offering suggestions for outcomes, or identifying patterns that might not be immediately apparent. Yet, fully autonomous decision-making without human oversight—a concept that could lead to AI-Driven Dispute Resolution (AIDR)—remains speculative and is not yet part of established ODR models. This paper will focus on AI's current capacity to optimise existing ODR processes, rather than exploring fully autonomous systems.

A particularly illustrative example of this is the development of Mediator Bot Harmony, an AI-powered tool created by Kleros to streamline the mediation phase of dispute resolution. This system uses OpenAI’s ChatGPT to facilitate both voice and text communication between disputing parties within a structured, human-like mediation. The model is hybrid in nature, combining automated interaction with the option of human oversight. It guides parties through a series of defined stages—from initial introductions and clarifications to the exchange and evaluation of proposals—with the aim of helping them reach a mutually acceptable solution before escalating the case to a jury. However, in instances where no agreement is reached, the unresolved issues are reformulated and submitted to the Kleros platform for adjudication. As in conventional mediation, this process remains non-binding, offering a more cost-effective and efficient AI-enhanced alternative (Dean & Ast 2024).

AI can be categorised into several types based on its applications: those that learn from the data provided and enhance their outputs over time, known as ML; those that assist with tasks in the physical world, such as robots; those that process and interpret visual information, such as facial recognition systems; and finally, AI specialised in human language understanding and processing, referred to as Natural Language Processing (NLP). Since ODR relies heavily on party input, the interpretation of their intentions, and the application of legal principles, NLP emerges as the most suitable AI for enhancing the efficiency of Kleros. For example, given the substantial amount of information presented to jurors, NLP could automate case summarisation, thereby streamlining submissions and aiding jurors in their decision-making. In instances where jurors lack legal expertise and may struggle with legal jargon, NLP could simplify complex terminology, making it more accessible. Regarding confidentiality and information retrieval—currently managed manually by Kleros staff—NLP could automate these tasks, offering a higher degree of privacy for the parties involved. Additionally, NLP could significantly improve the interpretation of legal intent, allowing jurors to better comprehend the true intentions behind the actions of the disputing parties.

This paper follows a problem-solving research approach and seeks to address the question: How can the integration of NLP improve the efficiency and decision-making processes within ODR systems like Kleros? The paper is divided into four main sections. The first section identifies the key challenges faced by current ODR systems, offering a general overview and highlighting their relevance to Kleros, along with the potential for addressing these challenges through the application of NLP. The following section provides a detailed analysis of NLP and its relevant techniques, emphasising their applicability within the ODR framework. Next, the paper demonstrates how NLP can be integrated into Kleros, exploring the potential benefits it could bring. Finally, the paper examines the possible drawbacks of implementing NLP, identifying areas that may require further exploration and discussion.

2. ODR Systems and Associated Challenges

ODR systems, for the purpose of this study, are defined as processes and tools developed by private entities that leverage the internet to facilitate the resolution of disputes between disputing parties, providing alternatives to traditional, state-run judicial mechanisms. It is crucial to note that the digital medium is an indispensable characteristic of ODR; any suggestion that ODR may merely involve technology as a supportive element, as Carneiro et al. (2014) contend, misrepresents its core nature. Carneiro et al.'s (2014, pp. 212-213) definition, which encompasses 'the use of these mechanisms in a technological context, either supported by technology or within a virtual computational environment,' lacks sufficient precision. This broader conceptualisation fails to recognise that, for a dispute resolution process to qualify as ODR, it must be conducted entirely in an online or digital environment. The term 'virtual computational environment' is also redundant, as the concept of conducting dispute resolution online inherently implies such an environment.

ADR systems typically involve three parties: the two disputing parties and an intermediary, such as a conciliator, mediator, or arbitrator, whose primary role is to facilitate the resolution of the dispute, either by identifying a mutually agreeable solution or by resolving the matter in accordance with the applicable law. In the context of ODR, however, some scholars argue that a “fourth” or even “fifth” party should be recognised (Carneiro et al., 2014, p. 214; Alessa, 2022, p. 324). Specifically, Katsh et al. (2001) describe technology as the “fourth party” in ODR, while Lodder (2006) introduces the concept of the technology provider as the “fifth party”. Nevertheless, we assert that ODR systems, much like traditional ADR systems, involve only three parties. The technology employed by ODR platforms is not autonomous in decision-making; instead, it primarily serves as a tool to gather and present relevant information for human decision-makers. In this sense, technology acts as a supporting instrument, not an active participant in the resolution process. For instance, in commercial disputes regarding product quality—such as disputes over whether a product is defective—ODR platforms typically use technology to collect and organise evidence, leaving the ultimate decision-making to a human intermediary.

Furthermore, the role of the intermediary in ODR systems is not singular; rather, it constitutes a combination of human involvement and technology as a tool. Technology, in this context, facilitates dispute resolution but does not function as an independent entity capable of making decisions. This observation holds true even in blockchain-based or smart contract-based ODR systems, such as Kleros, where jurors—not the technology itself—make decisions through majority voting. Thus, the process of dispute resolution in ODR can still be considered to involve only three parties: the two disputing parties and the human intermediary responsible for final decisions, as in traditional ADR systems.

Due to their similarities, ODR is often conflated with related concepts such as ADR, ADR in virtual environments, or virtual hearings in judicial courts (e.g. Alessa 2022, pp. 320-321, Lodder & Zelznikow 2005, p. 297). This misapplication of terminology can result in significant misunderstandings regarding the scope and nature of ODR, leading to the erroneous identification of problems that are not necessarily inherent to ODR systems.

For instance, within the context of ADR, aside from arbitration, methods such as mediation and conciliation often face the challenge of decision enforcement—an issue that distinguishes ADR from traditional judicial systems (Lodder & Zelznikow 2005, p. 296). In contrast, ODR systems, which are more akin to online arbitration, typically ensure the enforceability of decisions through specialised platforms, such as those used in consumer-seller disputes (Lodder & Zelznikow 2005, p. 298) or via mechanisms like smart contracts, as exemplified by Kleros.
Conversely, ADR in virtual environments tends to rely on third-party providers for conducting hearings, which may introduce complications such as unreliable internet connectivity, technical malfunctions, or restricted access to requisite platforms. ODR systems, while reliant on technology, are generally designed with a robust infrastructure capable of managing a wide range of disputes either automatically or with minimal human intervention, thus exhibiting greater resilience to minor technical issues.

Lastly, virtual courtrooms that seek to replicate the formality of traditional court procedures often encounter legal and procedural complexities that are ill-suited to online settings, such as challenges related to the admissibility of physical evidence or the logistics of remote witness testimony. ODR, by contrast, typically addresses less complex disputes and often circumvents the intricate rules of evidence and procedure that are customary in courtrooms, thereby enabling more efficient and streamlined resolution processes.

3. Leveraging AI in ODR

In the academic literature, there exists a misconception regarding AI, its abilities, and the consequent definitions. It is a common mistake to suggest that AI is a superhuman machine that acts more intelligently than humans or can learn in a progressively advanced manner with minimal initial input (Alessa 2022, p. 322). This confusion is likely caused by the existence of different types of AI, such as ML, which indeed continues to develop and learn with certain outputs, or robots that can make autonomous decisions. However, attributing the characteristics of different types of AI to the single concept of "AI" is an overgeneralisation that consequently leads to misconceptions in understanding AI. These discrepancies do not mean that AI cannot be defined without confusion, nor that a definition cannot be provided simply because AI is constantly evolving.

A growing international consensus on the definition of AI is beginning to emerge, with the Organisation for Economic Co-operation and Development (OECD) offering one of the most widely recognised definitions. This definition has also been adopted by the European Union (EU) in the formulation of its legal framework for AI. According to the OECD, an AI system is defined as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" (OECD 2019, "I. AGREES" definitions).

While the OECD's definition is both comprehensive and inclusive of various forms of AI, this paper proposes an alternative definition to address specific aspects relevant to its scope. This alternative definition underscores the capacity of AI systems to be trained, to learn, and to emulate human abilities. For the purposes of this analysis, AI is defined as a human-designed system or machine that is typically programmed and trained on an initial dataset and is capable of performing tasks that involve autonomous decision-making, data processing and interpretation, the replication of human physical or cognitive abilities, and iterative self-improvement through feedback or interactions.

Various approaches have been proposed for classifying AI within the context of ODR. Alessa (2022) suggests a classification that differentiates between supportive and substitutive systems. Supportive systems assist and influence human decision-making, while substitutive systems are designed to make decisions autonomously, replacing human judgment (pp. 326-330). This classification aligns with a broader distinction between non-autonomous and autonomous AI, which may be particularly relevant when disputing parties seek a fully autonomous ODR system to resolve their conflicts.

A more granular subcategorisation is provided for supportive AI systems, which Alessa further divides into decision support systems (DSS), knowledge support systems, and intelligent interface systems. DSS aim to "weigh up different factors and compute the optimal outcome"; knowledge support systems function as "intelligent search engines that present (or omit) relevant information in a comprehensible manner"; and intelligent interface systems leverage NLP. However, these categories broadly correspond to Expert Systems (ES), ML, and NLP in application-based classifications, which will be described in greater detail below.

In the case of substitutive systems, Alessa identifies two primary subcategories: case reasoning systems and rule-based systems. Case reasoning systems rely on knowledge of past outcomes to apply to current situations, combining features of both decision support and knowledge support systems. Rule-based systems, on the other hand, apply predefined principles and rules to a case, similarly integrating characteristics of both decision and knowledge support systems. This overlap raises questions about the necessity of distinguishing between the two.
While the classification of supportive and substitutive systems and their subcategories provides useful theoretical insights, it appears overly specific and somewhat redundant from a practical perspective. The distinctions between the subcategories lack sufficient differentiation to offer meaningful utility in real-world applications.

Carneiro et al. (2014, pp. 218-227) provide a more comprehensive classification of AI technology within the context of ODR, identifying eight primary categories. The first category, DSS, aligns with Alessa's (2022) initial classification. These systems are designed to facilitate the "generation and analysis of ideas... based on critical information, issuing substantiated recommendations, and compiling data that can inform the decision-making process." A significant distinction between the two frameworks is that Carneiro et al. characterise rule-based systems as a practical implementation of DSS, while Alessa categorises rule-based systems as substitutive rather than supportive. Furthermore, Carneiro et al. explicitly state that DSS are typically subject to human supervision and are not fully automated (p. 218). This further supports our earlier assertion that, at present, AI—specifically DSS—in ODR is, to varying degrees, consistently subject to human supervision.
The second category comprises ES, which emulate human expertise and knowledge in specific domains to make decisions based on predefined rules and knowledge bases (p. 219). ES are inherently static and can therefore be characterised as more rigidly rule-based than DSS, with limited capacity for adaptation unless explicitly updated through human intervention and subsequent analysis of outcomes. Whilst ES play a more active role in delivering solutions within established frameworks, they may also be encompassed within the broader classification of DSS. Furthermore, advancements in ML and NLP have increasingly blurred the traditional distinctions between DSS and ES, as modern ODR platforms often integrate elements of both. These platforms combine expert knowledge

bases with adaptive learning algorithms, thereby making a strict separation between these categories redundant.

The third system is the Knowledge-Based System (KBS), defined as “a collection of specialised facts, procedures, and judgment conventions” (p. 221). KBS is often confused with ES, as both focus on specialised knowledge; however, while ES function as AI advisers in specific domains by mimicking human expertise, KBS represent a database of specific knowledge. All ES fall within the category of KBS (Akerkar & Sajja 2009, p. 21), so it is more logical to present KBS as the overarching category and subsequently discuss ES as a subcategory.

The distinction between DSS and KBS is subtle and their functions may overlap. Both systems support decision-making: DSS rely on data analysis—with the user retaining ultimate control—whereas KBS utilise expert knowledge, often automating decisions or providing advice. Many modern systems integrate data analysis with expert knowledge to enhance decision-making. For instance, in healthcare, Clinical Decision Support Systems analyse patient data (in a DSS-like manner) while applying medical guidelines or rules (in a KBS-like fashion) to recommend treatments (Kalogeropoulos et al. 2003). Consequently, rather than maintaining a strict division between DSS and KBS or DSS and ES, it may be more rational to adopt a blended category.

Intelligent interfaces, defined by Carneiro et al. (2014, pp. 222–223) as a fourth category, do not constitute a distinct branch of AI. Rather, they represent an integrative application of multiple AI techniques designed to enhance user–computer interactions. These interfaces combine methods such as ML, NLP, and deep learning to offer advanced functionalities. Carneiro et al. primarily discuss their utility in tasks related to data organisation, compensating for incomplete or inaccurate user input, and filtering search results. Although these functions are closely tied to search mechanisms, the potential applications of intelligent interfaces in ODR extend far beyond simple data retrieval. For example, acting as virtual mediators Intelligent interfaces can leverage intelligent interfaces to facilitate negotiations by analysing communication for tone and emotion. Through sentiment analysis—a key NLP technique—the system can discern whether messages are conciliatory or confrontational, providing critical insights into the parties' intentions, which are often not detectable by conventional methods. Additionally, intelligent interfaces can support text summarisation and predictive analysis, further streamlining the resolution process by presenting complex legal information in an accessible and actionable format.

Since intelligent interfaces do not represent an independent technology but rather constitute an amalgamation of techniques such as ML, NLP, and deep learning, categorising them as a separate branch of AI is not entirely convincing.

Nonetheless, Alessa (2022, pp. 328–329) also classifies intelligent interfaces as supportive systems, placing them alongside DSS and knowledge support systems.

Case-Based Reasoning (CBR), or what Alessa refers to as case reasoning systems, constitutes the fifth category of AI applications in ODR (Carneiro et al. 2014, p. 225; Alessa 2022, p. 329). It functions as an advisory AI method that predicts outcomes by drawing analogies to previously decided cases. This technique closely mirrors the judicial reasoning process, particularly in common law systems or international adjudication, where judicial precedent serves as a complementary source for judicial rulings. The fundamental principle of CBR is to derive decisions based on past experiences rather than resolving each issue from first principles, thereby enhancing efficiency and consistency in legal reasoning. By identifying patterns and similarities with previous cases, CBR can streamline case analysis and decision-making. However, while this method optimises legal research and advisory processes, it does not fully replace human judicial reasoning.

A key limitation of CBR in law is its reliance on past cases without the ability to critically assess whether those precedents remain applicable in changing legal and societal contexts. Judges do not merely apply past rulings; they interpret, adapt, and sometimes depart from precedent based on evolving norms and values. Without mechanisms for such dynamic assessment, a purely CBR-driven system risks perpetuating outdated legal principles rather than ensuring justice that aligns with contemporary standards. Thus, while CBR can enhance legal decision-support systems, it requires integration with more advanced AI methods—such as ML for predictive analytics or intelligent interfaces for interactive legal reasoning—to ensure greater adaptability and contextual understanding.

With the seventh category, it becomes clearer that although Carneiro et al. (2014, p. 218) claim to classify the sub-fields of AI from the conflict resolution perspective, they ultimately present a compilation of AI methods and techniques, which, in themselves, do not constitute distinct AI types or categories. Therefore, the eighth category, legal ontologies, unlike the previously mentioned methods such as DSS or ES, is a technique used for structuring legal knowledge in a formal manner. For instance, in e-court systems, legal ontologies facilitate the classification and routing of cases to the appropriate adjudicators, while in legal research, they assist AI in identifying relevant case law and statutes based on structured legal terms. Importantly, legal ontologies are not strictly an AI technique but rather a broader concept within the field of knowledge representation. They are primarily based on knowledge representation and semantic technologies, rather than directly relying on ML or NLP, as is the case with some other AI methods.

The final, eighth category is Rule-based Systems (RBS), which, according to Alessa (2022, p. 330), falls under the category of substitutive systems. While case-based reasoning can be likened to the use of judicial precedent in common law systems, RBS are more comparable to the application of legal norms in civil law systems, where decisions are made by applying predefined rules under specific conditions. Although RBS and ES share similarities, particularly in their use of structured knowledge to make decisions, they differ in sophistication. Carneiro et al. (2014, p. 226) note that “these systems allow for ease of access to expert knowledge”, underscoring the capacity of RBS to replicate expert decision-making by systematically applying rules to specific cases, making them simpler and more domain-specific than ES.

Ultimately, the categorisation proposed by Carneiro et al. (2014), while more extensive than that of Alessa (2022), remains a classification of AI techniques and methods, rather than a comprehensive typology of AI. Additionally, the list of methods and techniques presented is not exhaustive, and with the rapid advancement of AI, it remains an open question whether a truly exhaustive typology can be constructed. The criteria for distinguishing between categories are not always clear. For instance, the decision to categorise legal ontologies separately, while not making similar distinctions for other techniques like ML or NLP, raises questions about the underlying logic of the classification. This uncertainty may partly stem from the fact that, when the paper was written, the authors themselves acknowledged that the application of AI in ODR was not widespread and was considered to be “playing a secondary role” (p. 230). Given the limited scope of AI's role at the time, the necessity of categorising methods and techniques becomes questionable.

Numerous classifications for AI can be derived in the context of ODR, extending beyond those based solely on techniques and methods. For example, based on AI capabilities, AI can be classified as narrow, general, or superintelligent. However, within the specific context of ODR, classifying AI according to its application—rather than its techniques, methods, or capabilities—provides a more effective means of understanding its role and impact. This application-based classification offers a clearer understanding of the technologies involved, the processes they govern, and their operational mechanisms, rather than merely characterising how a technology can be employed. Based on AI applications in ODR, three primary categories can be identified, which may be used individually or in combination: ML, ES, and NLP.

ML, which has been previously mentioned but not explicitly classified, refers to a branch of AI that enables systems to learn from data, identify patterns, and make decisions with minimal human intervention. In the context of ODR, ML can analyse historical cases and predict outcomes, functioning similarly to CBR under Carneiro et al. (2014) while also exhibiting characteristics of both substitutive and supportive systems as classified by Alessa (2022). However, ML possesses a unique ability to adapt and improve over time, refining its recommendations and decisions as new data becomes available. Moreover, ML can be trained using two primary approaches: supervised learning, where algorithms are trained on labelled data to predict outcomes, and unsupervised learning, where systems identify hidden patterns in data without prior labelling.

ES, referenced by both authors, replicate human expert judgment and assist in decision-making on specific issues. They integrate rule-based reasoning with knowledge representation techniques, such as semantic networks or frames, allowing the system to simulate legal reasoning by applying established rules to new situations. Unlike ML, which identifies patterns and adapts over time based on data, ES rely on predefined rules and do not improve or evolve without human intervention. Instead of leveraging statistical models to identify patterns, correlations, and trends, they depend on explicitly encoded human expertise.
NLP is focused on enabling machines to understand, interpret, and generate human language. In ODR, NLP can be employed to analyse legal documents, automate communication between parties, interpret user inputs (e.g., through chatbots or virtual assistants), and extract relevant information from large volumes of text. In theory, NLP can be combined with both ML and ES, but while ML makes NLP more adaptable and data-driven, ES typically use NLP to interpret human language and then apply rules or knowledge to derive conclusions
These three AI types were specifically selected for use in the ODR domain, though the scope of AI based on applications could be broadened to include robotics, computer vision, and autonomous systems. The crucial question of which AI technologies are most applicable to ODR is particularly relevant given the emerging nature of AI in this field, and the evolving relationship between AI and ODR. Alessa (2022) asserts that both AI and ODR have long histories, with AI "playing a role in the application of justice, preservation of rights, and the promotion of social values" (p. 323). However, this assertion appears to contrast with the current state of research, where AI is still being explored and tested in both public and private sectors. The EU AI Act highlights the regulatory challenges surrounding AI, while in the private sector, platforms like Kleros are just beginning to integrate AI into their ODR systems. Moreover, AI faces ongoing criticism for perpetuating and even amplifying biases in datasets, particularly in areas such as criminal justice (Barabas, 2020, p. 83). Instead of speculating about the potential for fully autonomous AI-driven ODR systems (Alessa, 2022, p. 325), this paper takes a pragmatic approach by focusing on the current use of AI as a tool within ODR.

All three— ML, ES, and NLP —can be integrated into ODR platforms like Kleros. However, this study will focus on identifying the most suitable option, as AI implementation demands specific expertise and significant resources, and the costs may not always justify the benefits. Prioritising the enhancement of Kleros' efficiency—particularly in supporting jurors in their tasks—is essential, and therefore, the chosen AI technology should directly address this need without introducing the added complexity of integrating multiple systems.

Kleros could relatively easily integrate an Expert System by developing a knowledge base of rules and guidelines based on legal norms and dispute resolution precedents. This would allow the creation of a system that makes decisions according to these fixed rules, ensuring predictable outcomes and simplifying development and deployment. However, Kleros' primary goal is not to replace the jury but to assist jurors in their decision-making process. Relying solely on an Expert System would be counterproductive, as it could limit the flexibility and judgment that human jurors bring to the platform. While such a system could automate the classification of disputes based on predefined rules, this approach would conflict with Kleros' model, where the parties involved choose the court in which to resolve their dispute, thus defining its categorisation. Moreover, the integration of an Expert System would require significant resources to build and maintain an extensive knowledge base (Rhim & Park 2019, p. 11), making it a costlier option compared to AI systems that better support jurors' tasks without replacing their role.

ML, on the other hand, requires a large dataset for training, relying on historical dispute data and continuous feedback from ongoing cases. While ML can enhance decision-making by predicting outcomes or offering recommendations based on past precedents, it introduces greater complexity compared to an Expert System. The model demands continuous training, testing, and updating, which makes it more resource-intensive. Moreover, Kleros' primary purpose is not to establish legal precedents but to provide swift and efficient dispute resolution across a wide range of case types—many of which may not be easily categorised. As a result, ML may not be the most suitable AI solution for Kleros. Although ML could potentially help parties by identifying trends in past cases and forecasting the likelihood of certain outcomes, its reliance on high-quality data and the need for ongoing fine-tuning make it inefficient for Kleros' core function, which emphasises the speed and simplicity of decision-making rather than complex predictive analytics.

Implementing NLP into Kleros could be a more suitable approach compared to ES or ML, as NLP directly supports the core functions of dispute resolution, particularly when it comes to managing and analysing textual data. NLP could assist Kleros by automating the extraction of relevant information from legal documents, streamlining case categorisation, and improving communication between parties, especially in multilingual settings. Although integrating NLP into Kleros would still require high-quality data to function effectively, it would need to be trained on a large corpus of legal texts to ensure accurate understanding and interpretation of the language used in disputes. Despite the initial costs, NLP could prove to be a more cost-effective option, as it would not require large datasets or continuous predictive model updates like ML, nor would it necessitate an extensive knowledge base to be manually curated, as with ES.

4. NLP Applications to Enhance Kleros

NLP, as defined by Chowdhary (2020, p. 604), is a collection of computational techniques designed for the automatic analysis and representation of human languages (both written and spoken), driven by theoretical foundations. These "theoretical foundations" refer to principles from linguistics (such as theories of syntax, or sentence structure, and semantics, or meaning (Nadkarni et al. 2011, p. 544)), computer science (including principles of data structures and algorithms), and AI (including neural networks and pattern recognition), which guide the development of algorithms and models for processing language data. The "automatic analysis" component refers to the ability of NLP systems to perform tasks like parsing sentences, identifying parts of speech, and extracting meaning from text without human intervention. For example, an NLP system can analyse the sentence "The claimant is entitled to compensation for breach of contract" by identifying "claimant" as a noun, "entitled" as a verb, and "breach of contract" as a legal concept, thereby breaking down the sentence into its grammatical and semantic components.

The "representation" aspect involves converting human language into structured formats that machines can process, such as vectors, graphs, or semantic networks. For instance, in legal document analysis, the sentence "The defendant shall pay damages amounting to €10,000" might be represented as a vector in a high-dimensional space, where its coordinates reflect key legal terms and monetary values. Similarly, in contract clause extraction, a sentence like "Party A shall deliver the goods within a reasonable time" could be represented as a graph with nodes for "Party A", "goods", and "reasonable time", connected by edges that capture the obligations and conditions (See Table 1).

Table 1: Graph Representation

Node 1	**Edge**	**Node 2**
Party A	Shall deliver	Goods
Party A	Shall deliver	Reasonable time
Goods	Within	Reasonable time

These capabilities are not merely theoretical but are already being applied in the aforementioned Kleros Harmony Mediator, where NLP is employed to interpret the conflict narratives submitted by the parties, identifying the relevant actors, actions,

and key legal or commercial terms. By extracting these elements and recognising commonly expected resolution patterns, the system can assist the parties in rephrasing their positions and formulating the conflict into two clearly defined resolution options—or binary outcomes—such as Outcome A: a refund, or Outcome B: a discount, in a commercial dispute, for example. Presenting the dispute in this binary format helps Kleros jurors, who vote on the most appropriate resolution based on the evidence provided (Dean & Ast, 2024).
NLP has made significant strides in recent years, however, when it comes to interpreting sentences and extracting meaningful information, the capabilities of these algorithms remain limited (Chowdhary, 2020, p. 604). This limitation is particularly evident in specialised domains such as law, where the interpretation of language often requires nuanced understanding and contextual awareness.

For instance, Holzenberger et al. (cited in Frankenreiter & Nyarko, 2022, p. 17) highlight that language models struggle to understand and apply legal rules from text without human guidance. Their experiment on tax calculations demonstrated that even when provided with relevant tax laws, the models performed poorly and showed no signs of improvement.

Chowdhary (2020, p. 605) highlights the inherent ambiguity of natural language as a fundamental challenge for NLP, since many words have multiple meanings or ambiguous parses (Nadkarni et al. 2011, p. 544), and the same sentence can often be interpreted differently depending on the context. For example, in the sentence “The defendant has to pay damages”, the word “damages” could refer to financial compensation for harm caused or to physical damage to property. Without proper context, an NLP system may struggle to discern the intended meaning, leading to potential errors in interpretation.

In addition, legal language often employs adaptable terms that introduce additional complexity for NLP systems (Frankenreiter & Nyarko 2022, p. 26). Phrases such as “The goods shall be delivered within a reasonable time”, “Party A shall use reasonable efforts to deliver the product”, or “Either party may terminate this agreement by providing reasonable notice” are common in contracts. These terms are intentionally flexible, allowing for adaptation to unforeseen circumstances without the need for exhaustive detail. In the event of a dispute, courts typically interpret terms like “reasonable” based on the facts of the case and the expectations of the parties involved. However, NLP systems may struggle to interpret such terms due to their inherent subjectivity and reliance on context.
Despite these fears, modern AI systems that employ NLP have become increasingly sophisticated and capable of recognising context. To illustrate this, consider the following scenario: A client hired a service provider to translate the sentence “Право на наследство оспаривается в суде” from Russian to English, with the condition that payment would be contingent upon the accuracy of the translation. The service provider submitted the translation: “The law of inheritance is being

disputed in court". The client, however, asserted that this translation was incorrect and refused to make payment, while the service provider maintained that the duties had been fulfilled. The crux of the issue lies in the Russian word "право", which can mean both "law" and "right" depending on the context. While the word "law" might seem plausible in certain contexts, advanced NLP models, such as GPT-4 and DeepSeek, concluded that the translation was inaccurate. These models suggested that a more accurate translation would be: "The right to inheritance is being disputed in court".

The capabilities of NLP are able to encompass both textual and spoken language analysis, and it is essential—following the approach of Trancoso et al. (2023, p. 26)—to distinguish between these two domains when discussing the abilities of NLP systems. According to Chowdhary (2020, p. 606), the key applications of NLP in the textual domain include: natural language translation, which enables the automatic conversion of text from one human language to another; information retrieval, which involves locating relevant documents or data within large corpora based on user queries; and information extraction, which focuses on identifying structured data such as entities, relationships, or events from unstructured text. NLP also supports text summarisation, the process of condensing lengthy documents into concise and coherent summaries; and question answering, where systems generate accurate responses based on textual inputs. In addition, topic modeling is employed to detect and categorise thematic structures within large datasets, while opinion mining—also known as sentiment analysis—identifies subjective content and emotional tone.

Chowdhary's categorisation primarily pertains to the textual side of NLP, and while the underlying tasks can be adapted for spoken input, such adaptation necessitates additional layers of technology, such as speech recognition and acoustic modelling. Tasks like information extraction and topic modelling are less directly applicable to raw speech, as they rely on textual structure and linguistic features—such as sentence boundaries, punctuation, and syntactic cues—that are either absent or difficult to detect in audio form (Piskorski & Yangarber 2012, p. 278). For example, information extraction is considerably more complex in speech due to the lack of visual markers like punctuation and the inherent variabilities of spoken language, such as hesitations, false starts, and informal grammar since many words have multiple meanings or ambiguous parses (Nadkarni et al. 2011, p. 544), and the same sentence can often be interpreted differently depending on the context. Similarly, topic modelling assumes a consistent textual representation, which is often compromised in spoken input due to factors such as background noise, accents, and speech disfluencies, all of which can degrade transcription quality. However, in theory, all other applications of NLP can also be extended to spoken language processing.

Notably, Piskorski and Yangarber (2012, p. 278) highlight the differing levels of accuracy in the perception of written and spoken language by NLP systems, with the former being more accurate than the latter. They provide an example demonstrating that even tasks like text summarisation are subject to recognition errors in spoken language, whereas such issues are not as prevalent with written text. Their proposal is grounded in the idea that humans rely on context to interpret information correctly. For instance, human language includes homonyms—words that are spelled and pronounced the same but have different meanings, such as “bank”, which can refer to a financial institution or the side of a river, or “light”, which can denote either illumination or a measurement of weight. Piskorski and Yangarber suggest that NLP systems should be provided with the same contextual understanding that humans use to reduce such recognition errors.

Although the study was conducted over a decade ago, insights can be drawn from contemporary personal experience with generative AI in textual formats. In such contexts, AI systems are generally capable of discerning the intended meaning and contextual nuances of a sentence, even when it contains spelling errors or is syntactically flawed. This stands in contrast to spoken language processing—such as that employed by systems like Amazon Alexa or Apple Siri—where even grammatically correct and clearly articulated speech may result in considerable recognition errors, underscoring the persistent challenges faced by speech recognition technologies.

This paper is limited to the analysis of NLP techniques applied to written, rather than spoken, language. The following sections will examine a range of NLP methods relevant to legal contexts, including data anonymisation, text summarisation, natural language translation, information retrieval, information extraction, question answering, topic modelling, opinion mining, and text and document classification, based on the classifications proposed by Chowdhary (2020, p. 606) and Trancoso et al. (2023, pp. 27-35). The only notable exclusion from this discussion is predictive modelling, as referenced by Trancoso et al. (2023, p. 34). This omission is intentional, given that predictive applications generally rely on large datasets of prior judicial decisions to produce reliable outputs. Not all international courts and tribunals possess a sufficiently extensive or accessible corpus of decisions to support such methods. For example, the International Court of Justice, having delivered fewer than 200 judgments to date (International Court of Justice, 2025), lacks the data volume required for effective implementation (Rhim and Park 2019, p. 22). In contrast, the European Court of Human Rights, with its tens of thousands of rulings, presents a more appropriate environment for predictive analytics (Rhim and Park 2019, p. 21). In the case of Kleros, which remains in a phase of ongoing development and wider adoption, predictive techniques appear less immediately relevant than other NLP approaches.

A) Anonymisation

The anonymisation of sensitive data is crucial for both traditional courts (Trancoso et al. 2023, p. 27) and ODR platforms such as Kleros. In legal proceedings, judges or jurors must review relevant documentation, which often contains sensitive or confidential information. In courts, such data must be protected against unauthorised disclosure. In the context of ODR platforms—particularly those based on blockchain—the need to safeguard sensitive content is even more pronounced, as anonymity is a foundational feature (Rabinovich-Einy 2002, p. 17, para. 45). Without the use of technologies such as NLP to automatically detect and anonymise such information, this task would need to be performed manually. This not only increases the operational burden but also calls into question the extent to which the process can genuinely be characterised as anonymous.

To implement anonymisation effectively in Kleros, NLP systems would need to be trained to identify and redact personally identifiable information (PII) and other contextually sensitive data within the submitted evidence and written arguments. This includes names, addresses, identification numbers, email addresses, phone numbers, car plates, bank account references, websites, and other data that could potentially reveal the identity of the parties involved (Trancoso et al. 2023, p. 27). State-of-the-art anonymisation tools use NER to detect entities. More advanced systems go further by recognising indirect identifiers that, when combined, might compromise anonymity (e.g. “the 40-year-old CEO of a Paris-based ODR startup”).

A generic anonymisation system typically consists of four interconnected modules (Trancoso et al. 2023, p. 27). The process begins with standardising the text by removing special characters, resolving abbreviations, and performing other necessary pre-processing tasks. The next stage involves a set of NER classifiers—statistical or neural models—trained to identify specific types of sensitive information. These classifiers often operate in parallel, each specialising in detecting a different entity type. The results from the classifiers are then passed to a voting or decision module, which aggregates the outputs and determines the most likely class for each identified sensitive entity. Finally, the anonymisation module replaces or masks the sensitive information by suppression, tagging, random substitution, or generalisation.

Applying all four anonymisation methods to the term “Kleros” would involve the following transformations: suppression would completely remove the term, replacing it with a placeholder such as [REMOVED]. Tagging would substitute “Kleros” with a more general label, like [ORGANISATION]. Random substitution would replace “Kleros” with an unrelated term, such as “Zenith”. Finally, generalisation would replace the specific name with a broader description, such as “Dispute Resolution Platform”.

The potential difficulty of implementation may stem from selecting the appropriate NER approach (Wen et al. 2020). Rule-based systems, while simple, may underperform when dealing with the unpredictable content submitted to Kleros. Statistical or ML-based models, on the other hand, require substantial amounts of annotated data (such as labels for persons, entities, etc.), which may be challenging to obtain (Frankenreiter & Nyarko 2022, pp. 23-24). Neural models, although highly effective, would require extensive training on legal documents. However, this issue will not be explored further in this paper.

B) Text Summarisation

Text summarisation is designed to reduce lengthy documents into more concise summaries while retaining the key information (Trancoso et al. 2023, p. 31). The result or the output of summarisation is markedly different from that of information retrieval, as summarisation generates a condensed version of the input text. While information retrieval offers access to relevant content, summarisation enables quicker comprehension of that content by distilling it to its most important elements.

Text summarisation, based on output type, can be classified into two main approaches: extractive summarisation and abstractive summarisation (Trancoso et al. 2023, p. 31; Rahimi et al. 2017, p. 56). It is important to note, however, that this classification is not exhaustive. Other classification methods—based on factors such as level of detail, content type, limitations, number of input texts, and language support—are also provided by Rahimi et al. (2017, pp. 56–57). However, this section will focus exclusively on extractive and abstractive summarisation. Extractive summarisation works by identifying and selecting the most relevant sentences from the original text (Rahimi et al. 2017, p. 0056). This technique typically ranks sentences based on specific criteria such as frequency, importance, or relevance to the main theme. The chosen segments are then stitched together to form a summary. In contrast, abstractive summarisation involves generating entirely new sentences that paraphrase the original content (Rahimi et al. 2017, p. 0056). This method requires a more sophisticated understanding of the text. The technical processes behind both extractive and abstractive summarisation typically involve several shared steps, including the aforementioned text preprocessing, tokenisation, and vectorisation (Chai 2023).

Abstractive summarisation relies on advanced ML models and can be computationally intensive. The time and resources required—such as real-world knowledge and semantic class analysis—to process and summarise large volumes of data may become prohibitively expensive, both in terms of computational infrastructure and operational expenditure (Suleiman & Awajan 2020, p. 2). Nevertheless, for platforms such as Kleros, which are likely to handle a relatively lower volume of submissions in real time, the use of abstractive summarisation could still prove beneficial. Within the Kleros, such summarisation could be used to generate brief summaries of legal disputes, evidence, or case law, allowing jurors to perform more efficiently.

Additionally, summarisation can be utilised to produce standardised "juror briefs"—concise, consistently formatted documents that typically include a clear and objective explanation of the dispute, outlining the relevant context and the issues at stake. Such explanations are expected to maintain neutrality, focusing on factual information rather than assigning blame. By streamlining the case preparation process in this manner, the summarisation of disputes has the potential to enhance juror efficiency and reduce operational costs, as highlighted by Dean and Ast (2023).

C) Natural Language Translation

Natural language translation is a subfield of computational linguistics concerned with the automatic rendering of text or speech from one language into another. It constitutes a critical application within the broader domain of NLP (Hirschberg & Manning 2015, p. 261), particularly in multilingual legal contexts where precise translation is indispensable for ensuring fairness and mutual understanding. In the context of international dispute resolution platforms such as Kleros, natural language translation can play a key role in bridging linguistic divides between parties from diverse cultural and linguistic backgrounds, thereby enhancing the access to justice.

The process of natural language translation typically comprises several stages. Initially, the input—whether text or speech—is subject to pre-processing in order to standardise and normalise the linguistic features of the source language. This stage may involve tasks such as tokenisation (the process of breaking down text into smaller units), part-of-speech tagging (labelling each word in a sentence with its grammatical category), or syntactic parsing (analysing the grammatical structure of a sentence to identify relationships between words) (Hirschberg & Manning 2015, pp. 261-262; Chai 2023). Following this, the translation model is applied to generate the corresponding output in the target language. The underlying models used for translation can range from earlier rule-based or statistical approaches to the now predominant neural machine translation (NMT) systems (Bahdanau et al. 2014, p. 1; Kalchbrenner & Blunsom 2013). NMT has emerged as the state-of-the-art technique owing to its ability to capture contextual relationships between words, phrases, and sentences, thereby producing more fluent and semantically accurate translations (Kalchbrenner & Blunsom 2013, p. 1700).

Neural machine translation models are trained on extensive corpora of parallel texts (Bahdanau et al. 2014, p. 4), enabling them to learn the complex syntactic and semantic relationships that underpin language use. These models represent language in high-dimensional vector spaces, allowing them to retain contextual meaning and adapt to varying linguistic patterns (Bahdanau et al. 2014, pp. 1-2).

When trained on domain-specific corpora, such as legal texts, these systems can further enhance their performance by learning specialised terminology, formal structures, and interpretive conventions characteristic of legal discourse. For platforms like Kleros, where disputes are adjudicated across jurisdictions and languages, the integration of real-time, domain-sensitive translation could significantly enhance the accessibility for both parties and jurors.

The primary challenge when dealing with multiple languages in the legal context lies in the inherent complexity of legal language, which is formal, technical, and highly context-dependent (Frankenreiter & Nyarko 2022, p. 26). Such language requires precise terminology that can lack direct equivalents across different languages. In legal contexts, even minor differences in phrasing can carry significant implications. Moreover, legal translation demands not only linguistic accuracy but also a deep understanding of differing legal traditions and associated terminology—an aspect that current machine translation systems often struggle to accommodate.

The risks of misinterpretation in legal translation are well-documented, particularly in international law, where multilingual treaties serve as a primary source of law (Rhim and Park 2019, p. 26). One of the most notable historical examples is the Treaty of Waitangi in New Zealand, where a mistranslation of the word "sovereignty" in the Māori version as compared to the English version gave rise to enduring legal and political controversy. Similarly, the Treaty of Wuchale between Italy and Ethiopia was the subject of significant misunderstanding due to divergent translations, ultimately contributing to the outbreak of conflict (Masiola et al. 2015, pp. 86-91).

Equally important is the question of legal expertise. Platforms like Kleros often rely on jurors who may lack formal legal training or consistent familiarity with diverse legal traditions. This variability in background legal knowledge among participants from different linguistic and cultural contexts can affect the quality of dispute resolution. Consequently, the success of such platforms depends not only on the ability of AI tools to accurately translate and interpret the core ideas of a case but also on bridging gaps in legal knowledge among participants from varied linguistic and cultural backgrounds.

D) Information Retrieval

The bureaucratic nature of legal systems often entails the use of formalised and excessively lengthy documents. To mitigate the need for a time-consuming, page-by-page review in search of relevant information, the application of information retrieval proves invaluable—for instance, in extracting documents or pieces of text. The key takeaway is that information retrieval, in a legal context, can be effectively applied to both evidence analysis and legal research. In terms of evidence analysis, it can enable jurors to swiftly identify relevant pieces of evidence from extensive volumes of submitted material. Simultaneously, it can facilitate the search of legal databases for pertinent statutes, case law, or precedents that address the specific legal issue at hand (Sansone & Sperlí 2022, p. 10).

Information retrieval can play an important role in prediction and summarisation. In prediction, the process is relatively straightforward, as retrieving analogous information helps identify patterns and forecast potential outcomes. In summarisation, the retrieved information is organised and condensed into a concise format, distilling the key points and facilitating more efficient navigation through complex legal texts.

The process of information retrieval begins with text pre-processing, which typically includes tokenisation, stemming (reducing words to their base or root form), lemmatisation (ensuring words are in their proper dictionary form), and the removal of stop words (Jabbar et al. 2023, p. 133684; Chai 2023). These steps serve to standardise language input and reduce linguistic variation. Following this, both documents and user queries are transformed into numerical representations—typically through vectorisation techniques—which enable the computation of similarity scores (Jabbar et al. 2023, p. 133684). These scores are then used to retrieve content deemed most relevant to the user's query (Jabbar et al. 2023, p. 76593).

At a technical level, the core components of an information retrieval system—namely pre-processing, vectorisation, and similarity scoring—can be relatively easily integrated into platforms such as Kleros. Indeed, many existing platforms already make use of such systems. For instance, Google Scholar relies on information retrieval to index academic content, facilitate search, and support citation analysis.

While implementing this technique into Kleros may be beneficial, it should be combined with other methods—such as text classification or natural language translation—to address the potential need to accommodate multiple languages, the unstructured and inconsistently formatted nature of submissions, the presence of colloquial expressions or incomplete information, and—most importantly—the requirement for annotated training data.

E) Information Extraction

Unlike information retrieval, information extraction seeks to identify and structure specific pieces of information from unstructured or semi-structured text (Trancoso et al. 2023, p. 31). While information retrieval focuses on finding and returning entire documents or text segments relevant to a query (e.g., returning contracts that mention “breach of supply agreement”), information extraction delves into the content of those documents to extract concrete facts, relationships, or entities (e.g., identifying the parties to the contract, the specific clause breached, the amount in dispute, and the governing law).

The process of information extraction typically comprises several core components. The first is Named Entity Recognition (NER), which involves detecting and classifying entities such as people, organisations, locations, legal instruments, or dates (Pudasaini et al. 2021, pp. 700-702; Singh 2018, p. 2). This is followed by relation extraction, which identifies and categorises semantic relationships between entities (Singh 2018, p. 2)—such as the connection between parties to the dispute, or the legal grounds cited in a claim. In more advanced implementations, event extraction (Singh 2018, p. 4) may also be used to pinpoint specific occurrences described in the text, such as the signing of a contract.

Implementing information extraction in Kleros is considerably more complex than information retrieval, as it requires not only locating relevant content but also understanding and structuring it. While nformation retrieval systems can be integrated relatively easily, information extraction demands high-quality annotated training data and sophisticated natural language processing models, particularly for tasks like named entity recognition and relation extraction.

F) Text Classification

Text classification, mentioned by Trancoso et al. (2023, p. 29) though not listed by Chowdhary (2020), aligns to some extent with information extraction—where it may serve as an intermediate step to identify specific categories of information—and with opinion mining, which classifies text based on sentiment or subjective content. However, text classification is not necessarily dependent on these contexts and should be recognised as a distinct category of NLP tasks in its own right.

Text classification—the process of labelling or categorising textual data—serves to organise content and may be applied to individual sentences, paragraphs, or entire documents (Dogra et al. 2022, p. 3). Document classification, followed by post-text classification, could prove highly useful in the context of Kleros, particularly for automatically identifying the nature of a case, filtering out irrelevant submissions, prioritising key pieces of evidence, or routing disputes to jurors with the relevant expertise. At present, parties are tasked with selecting the appropriate court themselves, a step that assumes a degree of legal knowledge and does not preclude the possibility of error. In this regard, automatic classification may offer an alternative or additional support. This function could also assist in sorting already submitted documents, especially evidence, thereby easing the jurors' workload and allowing them to concentrate on the most salient materials.

Text classification systems typically rely on supervised ML methods, wherein algorithms are trained on annotated datasets containing examples of texts paired with their corresponding labels (Dogra et al. 2022, p. 2) —such as "contract dispute," “defamation,” or “intellectual property infringement.” Through this training, the system learns to associate patterns in language—ranging from vocabulary and syntax to broader semantic cues—with predefined categories. Once trained, the model is able to generalise these associations and apply them to classify other documents.

More advanced techniques draw on deep learning architectures, such as convolutional neural networks (CNNs) or transformer-based models like BERT (Dogra et al. 2022, p. 15, 17), which are capable of capturing complex contextual dependencies and subtle semantic features across longer spans of text. Such models tend to perform particularly well in legal and dispute resolution contexts (Trancoso et al. 2023, p. 29), where classification depends not merely on surface-level indicators but on a deeper grasp of argumentative structure, framing, and tone.

In practice, implementing text classification in Kleros would require the compilation of a representative and sufficiently diverse training corpus reflecting the variety of cases typically submitted to the platform. This, however, may present a challenge—not only for Kleros but for many dispute resolution bodies that do not possess an extensive, annotated dataset from which to train such models (Frankenreiter & Nyarko 2022, pp. 23-24).

G) Question Answering

Question answering is a subfield of NLP is concerned with the development of systems that can automatically provide accurate and contextually appropriate responses to questions posed in natural language (Ferret et al. 2002, p. 136). Unlike traditional information retrieval systems, question answering systems are designed to deliver precise answers, thereby significantly reducing the user’s cognitive load.

The architecture of a question answering system typically builds upon and integrates other NLP techniques, particularly information retrieval and information extraction (Chowdhary, 2020, p. 634). Information retrieval is employed to locate and rank relevant documents or text segments from a broader corpus based on the user’s query. Once a relevant subset of data is identified, information extraction techniques are applied to isolate pertinent entities, events, or relations. The final stage depends on whether the system follows an extractive or abstractive approach. Extractive question answering selects the most relevant span of text within the source material. In contrast, abstractive question answering leverages generative models to synthesise new responses by drawing on multiple inputs, thereby providing answers in more natural and concise language.

Since question answering primarily serves as an automated method for responding to procedural and administrative questions, it may appear less pertinent to Kleros than other NLP techniques, given that the platform is designed to remain accessible to the general public. Its implementation could still prove valuable in enhancing the usability of the platform for both parties and jurors. This is particularly relevant in relation to aspects involving the Kleros token, pinakion, such as its acquisition and the process of staking it in disputes. Thus, a question answering system could simplify user interactions on token-related operations, thereby reducing entry barriers.

H) Topic Modeling

Topic modeling is an unsupervised ML technique used to discover hidden thematic structures within large collections of text. By grouping related words into topics, this method helps to uncover overarching themes that may not be immediately apparent (Snyder 2015, p. 86).

Topic modeling can be used for document clustering (Snyder 2015, p. 90), where large volumes of legal documents are grouped based on common themes, making it easier for jurors to navigate. For instance, in Kleros, topic modeling could assist in organising past cases according to relevant issues. Alternatively, it can be used for issue identification by automatically detecting the key issues within a set of documents, such as analysing past legal decisions. This method can help categorise disputes, identify relevant precedents, or even group evidence under similar headings.

The most common technique for topic modeling is Latent Dirichlet Allocation (LDA), a probabilistic model that assumes each document is a mixture of various topics and that each topic is a mixture of words (Snyder 2015, pp. 94-95). LDA can discover these topics by analysing word co-occurrence patterns across documents and inferring a set of topics that best explains the observed data. LDA, along with other techniques like Non-Negative Matrix Factorisation (NMF) or Latent Semantic Analysis (LSA), can be particularly valuable when documents are extensive, yet the key issues are subtle or diffuse.

An important consideration is the dynamic nature of the legal system, as well as human language in general, where interpretations of legal concepts evolve over time. While topic modeling can be effective in the short term, it may require periodic updates to maintain relevance and accuracy in the long term.

I) Opinion Mining

Opinion mining, also known as sentiment analysis, focuses on determining the sentiment or opinion expressed in a piece of text. This technique aims to classify text as expressing positive, negative, or neutral sentiment, or, in more advanced implementations, to detect more nuanced emotions such as anger, joy, or sadness (Yadollahi et al. 2017, 25:2). In legal contexts, opinion mining can be particularly useful for analysing both legal documents and user-generated content, such as public opinions, court rulings, or social media discussions surrounding legal cases (Eliot 2020).

The process of opinion mining typically starts with text pre-processing, which includes tokenisation, stop word removal, and stemming (Chai 2023). After pre-processing, sentiment classification models are applied to the text to predict the sentiment expressed within it. These models can be rule-based, relying on predefined lists of positive and negative words, or ML-based, which require training on labelled data to identify patterns and features indicative of sentiment (Sun et al. 2017, pp. 11, 15).

For platforms like Kleros, opinion mining could also assist in the analysis of evidence presented by parties, revealing the tone or sentiment underlying the language of submissions, since understanding the underlying intentions or opinions of the parties to a contract is crucial in legal practice. Intentions can often provide insight into the true meaning of contract terms, helping to resolve ambiguities or disputes. For example, if a party's statements or actions suggest a certain understanding or intention behind a clause, this can help clarify the interpretation of contractual obligations or the expectations of the parties involved.

In contract law, the intention of the parties is foundational to determining the existence and scope of contractual obligations (Kraus & Scott 2009, p. 1046). Knowing the sentiment or opinion of parties helps interpret their actions, offers, and agreements in light of their motivations and goals. This is particularly important in cases of ambiguity, disputes over performance, or when the plain language of a contract does not adequately capture the underlying objectives.
It is important to note that sentiment analysis in legal language can be particularly challenging due to the formal and technical nature of legal writing, as sentiment may not always be explicitly stated. For instance, a court ruling may appear neutral on the surface but convey underlying tones of dissent or approval when considered within a broader legal context.

5. Further Discussion

This research aimed at identifying the primary techniques of NLP that would bring the most significant benefit to ODR platforms, specifically to Kleros. We aimed to explain why the focus is on NLP rather than on ML or ES, as well as to highlight the many-sided aspects of NLP and its application to texts—what it can do and which aspects should be paid special attention to. Some of these aspects were already mentioned in the introductory sections and throughout the text, such as the limited sophistication and ongoing debates surrounding NLP, problems inherent to natural language due to multiple meanings, legal terminology or hidden intentions, the challenge of handling multiple languages, inconsistencies in documentation, high demands on computational infrastructure and operational expenditure, lack of a diverse and representative training corpus, and so on.

There are a few additional points — or points for further investigation — that were found in the academic literature but will not be discussed in this paper. Instead, they are briefly mentioned in this concluding section to motivate others to conduct further research.

One critical technical challenge highlighted by Frankenreiter and Nyarko (2022, p. 21) is the difficulty that modern language models face when processing long legal documents. While they describe this as exponential complexity, it is more accurate to refer to the quadratic complexity typical of most transformer-based models—whereby doubling the input length results in a fourfold increase in processing time. This technical limitation poses a serious challenge in legal contexts, where documents frequently exceed dozens of pages and require an understanding of extended textual dependencies.

To cope with this, a common workaround involves splitting documents into smaller sections. However, this method often compromises the coherence of legal texts, where the meaning of one clause may depend on earlier provisions or contextual elements spread across the document. As Frankenreiter and Nyarko (2022, pp. 21–22) argue, segmenting texts in this way can lead to incomplete or even misleading interpretations.

Compounding these technical limitations is the problem of bias in NLP models trained on narrow or non-representative datasets. In legal applications—where impartiality is paramount—such biases risk perpetuating inequalities or disadvantaging specific parties (Frankenreiter & Nyarko, 2022, pp. 24–25). This issue is exacerbated by the reluctance of legal institutions to share representative documents for training, partly due to commercial incentives to maintain exclusive control over legal drafting and interpretation (Frankenreiter & Nyarko, 2022, p. 26).

Given these data constraints, transfer learning represents a promising direction for future research. A key question is whether models trained on data from other legal or consumer dispute resolution contexts can be adapted for use within Kleros, potentially alleviating the limitations caused by the scarcity of Kleros-specific data. Tasks such as anonymisation, text classification, topic modelling, and opinion mining are particularly well-suited to cross-domain training, as they rely on generalisable linguistic patterns. In contrast, moderately complex tasks—such as summarisation, information retrieval, and question answering—may require more extensive domain-specific adaptation to capture the nuances of Kleros disputes effectively. The most difficult tasks to transfer are information extraction and legal translation, which would likely necessitate fine-tuning and empirical evaluation to assess their suitability and effectiveness within the Kleros platform.

To make these insights actionable, Kleros should adopt a phased implementation strategy that balances feasibility and impact. An initial phase could prioritise foundational NLP applications, such as extractive text summarisation and rule-based anonymisation, to address immediate needs like reducing juror workload and ensuring data privacy. Subsequent phases might introduce operational tools—such as multilingual translation and case classification—and later advanced capabilities like sentiment analysis and information extraction. These developments would build progressively on Kleros's existing infrastructure, including the Harmony mediator.

Finally, the scope of NLP applications in ODR can be expanded beyond dispute analysis to include administrative and preparatory functions. Dimitropoulos (2023, para. 15) observes that AI can be used by parties to strengthen their arguments during both treaty negotiations and court hearings, a point relevant not only to international courts but also to Kleros-like platforms. For instance, Kleros Enterprise—a specialised unit within Kleros that supports parties in submitting disputes without requiring direct interaction with blockchain or cryptocurrency, with the Kleros team managing all protocol processes such as dispute submission, juror selection, and case management (Ast et al., 2024)—employs NLP to enhance the structure of parties' arguments. Clients are then offered the choice to adopt the improved argument or retain their original submission. This example suggests that the scope of NLP applications in ODR can be broadened to include support for argument refinement and related bureaucratic functions.
These few points highlight both challenges and opportunities—such as transfer learning and argument enhancement—that have the potential to significantly improve Kleros' dispute resolution processes and user experience, underscoring the need for further research.

6. Conclusion

Coming to the conclusion of this paper and addressing the primary research question — How can the integration of NLP improve the efficiency and decision-making processes within ODR systems like Kleros? — it is possible to identify ten core NLP techniques that can be effectively applied to written data, including texts and legal documents. These are: anonymisation, text summarisation, natural language translation, information retrieval, information extraction, text classification, question answering, topic modelling, and opinion mining. This list is not exhaustive; for example, additional applications such as support for position preparation and argument enhancement also show promise, as illustrated by Kleros Enterprise.

Each of these techniques is designed to process large volumes of legal or factual text and generate outputs that can enhance the understanding of factual claims, legal rules, or even underlying party intentions. For instance, text classification and information retrieval can support the rapid sorting and identification of relevant claims or precedents; anonymisation ensures privacy and compliance with legal standards; and summarisation and translation facilitate juror access to complex or multilingual material. Some of these techniques — such as text classification or information retrieval — are relatively easier to implement on smaller or resource-limited ODR platforms due to their lower computational and data requirements. Others, such as opinion mining or question answering, may demand more sophisticated infrastructure and access to high-quality training data.

It is important to emphasise that many of these methods are interconnected and not always used in isolation. For example, question answering often builds on information extraction and text classification, while topic modelling may support summarisation and opinion mining by identifying recurring themes or sentiments across the corpus. In practice, the application of one technique often triggers the need for another, forming an integrated NLP pipeline.

The purpose of this research was not to provide an exhaustive implementation model, but rather to identify the most promising NLP techniques for future investigation and eventual application within platforms like Kleros.

Bibliography

1. Akerkar, R., & Sajja, P. (2009). Knowledge-based systems. Jones & Bartlett Publishers.

2. Alessa, H. (2022). The role of Artificial Intelligence in Online Dispute Resolution: A brief and critical overview. Information & Communications Technology Law, 31(3), 319-342.

3. Ast, F., Pernas, M., & T, F. (2024). Kleros Enterprise: Dispute resolution for companies and governments. Kleros Blog. Retrieved June 30, 2025, https://blog.kleros.io/kleros-enterprise/

4. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

5. Barabas, C. (2020). Beyond Bias: Re-imagining the Terms of "Ethical AI" in Criminal Law. Georgetown Journal of Law & Modern Critical Race Perspectives, 12(2), 83-221.

6. Carneiro, D., Novais, P., Andrade, F., Zeleznikow, J., & Neves, J. (2014). Online dispute resolution: an artificial intelligence perspective. Artificial Intelligence Review, 41, 211-240.

7. Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

8. Chowdhary, K. R. (2020). Fundamentals of artificial intelligence (1st ed.). Springer New Delhi. https://doi.org/10.1007/978-81-322-3972-7

9. Dean, R., & Ast, F. (2023). Kleros Mediation Bridge: A Cohesive Approach Blending Traditional Mediation and Kleros Blockchain Arbitration. Kleros Blog. Retrieved June 30, 2025, from https://blog.kleros.io/innovating-dispute-resolution-a-cohesive-approach-blending-traditional-mediation-and-kleros-blockchain-arbitration/

10. Dean, R., & Ast, F. (2024). Harmony, the Kleros Mediator Bot. Kleros Blog. Retrieved June 27, 2025, from https://blog.kleros.io/harmony-the-kleros-mediator-bot/

11. Dimitropoulos, G. (2023). Artificial intelligence and international adjudication. In Max Planck Encyclopedia of Public International Law. Oxford University Press. https://doi.org/10.1093/law-mpeipro/e3888.013.3888

12. Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art NLP models. Computational Intelligence and Neuroscience, 2022(1), 1883698.

13. Eliot, L. (2020). Legal Sentiment Analysis and Opinion Mining (LSAOM): Assimilating Advances in Autonomous AI Legal Reasoning. arXiv preprint arXiv:2010.02726.

14. Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., Monceaux, L., ... & Vilnat, A. (2002). How NLP can improve question answering. KO Knowledge Organization, 29(3-4), 135-155.

15. Frankenreiter, J., & Nyarko, J. (2022). Natural language processing in legal tech. In D. Engstrom (Ed.), Legal tech and the future of civil justice (forthcoming). SSRN. https://ssrn.com/abstract=4027030 or http://dx.doi.org/10.2139/ssrn.4027030

16. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

17. International Court of Justice. (2025). List of all cases. Retrieved on May 14, 2025 from https://www.icj-cij.org/list-of-all-cases.

18. Jabbar, A., Iqbal, S., Tamimy, M. I., Rehman, A., Bahaj, S. A., & Saba, T. (2023). An analytical analysis of text stemming methodologies in information retrieval and natural language processing systems. IEEE Access, 11, 133681-133702.

19. Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1700-1709).

20. Kalogeropoulos, D. A., Carson, E. R., & Collinson, P. O. (2003). Towards knowledge-based systems in clinical practice: Development of an integrated clinical information and knowledge management support system. Computer methods and programs in biomedicine, 72(1), 65-80.

21. Katsh, E. E., Katsh, M. E., & Rifkin, J. (2001). Online dispute resolution: Resolving conflicts in cyberspace. John Wiley & Sons, Inc.

22. Kraus, J. S., & Scott, R. E. (2009). Contract design and the structure of contractual intent. NYUL Rev., 84, 1023.

23. Lodder, A. R. (2006). The Third Party and Beyond. An analysis of the different parties, in particular The Fifth, involved in online dispute resolution. Information & Communications Technology Law, 15(2), 143-155.

24. Lodder, A. R., & Zelznikow, John. (2005). Developing an Online Dispute Resolution Environment: Dialogue Tools and Negotiation Support Systems in Three-Step Model. Harvard Negotiation Law Review, 10, 287-338.

25. Masiola, R., Tomei, R., Masiola, R., & Tomei, R. (2015). Manipulating Treaties. Law, Language and Translation: From Concepts to Conflicts, 73-94.

26. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

27. Organisation for Economic Co-operation and Development. (2019). Recommendation of the Council on Artificial Intelligence (OECD Legal Instrument No. OECD/LEGAL/0449). Retrieved January 11, 2025, from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

28. Piskorski, J., & Yangarber, R. (2012). Information extraction: Past, present and future. In Multi-source, multilingual information extraction and summarization (pp. 23-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

29. Pudasaini, S., Shakya, S., Lamichhane, S., Adhikari, S., Tamang, A., & Adhikari, S. (2021). Application of NLP for information extraction from unstructured documents. In Expert clouds and applications: Proceedings of ICOECA 2021 (pp. 695-704). Singapore: Springer Singapore.

30. Rabinovich-Einy, O. (2002). Going public: Diminishing privacy in dispute resolution in the Internet age. Va. JL & Tech., 7, 1.

31. Rahimi, S. R., Mozhdehi, A. T., & Abdolahi, M. (2017). An overview on extractive text summarization. In 2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI) (pp. 0054-0062). IEEE.

32. Rhim, Y. Y., & Park, K. (2019). The applicability of artificial intelligence in international law. Journal of East Asian and International Law, 12, 7.

33. Sansone, C., & Sperlí, G. (2022). Legal information retrieval systems: state-of-the-art and open issues. Information Systems, 106, 101967.

34. Singh, S. (2018). Natural language processing for information extraction. arXiv preprint arXiv:1807.02383.

35. Snyder, R. M. (2015). An Introduction to Topic Modeling as an Unsupervised Machine Learning Way to Organize Text Information. Association Supporting Computer Users in Education.

36. Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. Mathematical problems in engineering, 2020(1), 9365340.

37. Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information fusion, 36, 10-25.

38. Trancoso, I., Mamede, N., Martins, B., Pinto, H. S., & Ribeiro, R. (2023). The Impact of Language Technologies in the Legal Domain. In Multidisciplinary Perspectives on Artificial Intelligence and the Law (pp. 25-46). Cham: Springer International Publishing.

39. Wen, Y., Fan, C., Chen, G., Chen, X., & Chen, M. (2020). A survey on named entity recognition. In Communications, Signal Processing, and Systems: Proceedings of the 8th International Conference on Communications, Signal Processing, and Systems 8th (pp. 1803-1810). Springer Singapore.

40. Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2), 1-33.