Bridging Machine-Readable Code of Regulations and its Application on Generative AI: A Survey

Samira Yeasmin, Bader Alshemaimri
Software Engineering Department, King Saud University, Riyadh, Saudi Arabia

***Abstract*—Machine-Readable Code (MRC) and Machine-Readable Regulations (MRR) enable the conversion of complex regulations into structured formats such as JSON, XML, and X2RL, allowing machines to parse and interpret regulatory texts efficiently. Currently, organizations face challenges in regulatory compliance due to the complexity of regulations, frequent updates, and difficulty in identifying changes that impact policies and procedures. Existing literature provides guidance to a certain extent on how to anticipate regulatory modifications or ensure timely compliance. This review examines current literature on applying machine learning (ML) and Generative AI (GenAI) to extract, structure, and interpret regulatory content. It surveys techniques for converting regulations into machine-readable formats, predicting regulatory changes, and assessing alignment with real-world modifications issued by regulatory bodies. The findings indicate that using MRC, MRR, and AI enables automated compliance checks, faster detection of violations or errors, standardized compliance processes, real-time monitoring, and automatic report generation. These approaches can significantly enhance regulatory adherence across industries, particularly in sectors such as finance, where compliance is critical.**

***Keywords*—Regulatory compliance; natural language processing; machine learning; machine-readable code; Machine-Readable Regulations; generative AI; large language models; RegTech; conflicting regulations; regulation issuance**

I. Introduction

While regulations are an essential part of modern society, it is important to follow the process of regulatory change management. Regulations help to protect people and the environment, but they can be complex and difficult to understand. The complexity of regulations makes it challenging for businesses and individuals to comply with; therefore, recent studies have shown a vast amount of contribution to this field of science, as discussed in section 4. One way of ensuring regulatory compliance is to convert the regulations into machine-readable code (MRC) that can be easily understood and processed by computers.

Once converted, they can be used to train Large Language Models (LLMs) to understand and respond to them. LLMs work with massive amounts of text data and have the capability to understand and generate human language. On the basis of a given input, it has the ability to predict the likelihood of word sequences or generate new text. The larger the dataset to train this neural network, the more accurate the outcomes will be. It can be trained on large datasets of text, such as laws, regulations, and other legal documents. Given the capabilities of LLMs, an AI tool can be developed to highlight similarities between new policies and outdated regulations to help people understand the areas of change. The tool could also be used to identify the entity or organization that issued each regulation, which can help understand the context of the regulation and its applicability to a particular situation. This can allow the regulatory industry to check for compliance with regulations automatically.

Developing a tool based on fine-tuned LLMs to highlight similarities between regulations and updated policies to identify areas of change for the issuing entity or organization is a challenging task, but it has the potential to be a valuable tool for businesses and individuals who need to comply with regulations. Given that the process of regulatory change management requires a significant amount of time because of the complexity of the regulations, the long regulations, and the amount of text that requires investigation to identify change requires a significant amount of manual work and effort. A generative AI can generate new and meaningful content, including text, images, and audio, based on a huge amount of training data. With that being said and noting the capabilities of GenAI, a GenAI-based tool can make a significant impact on regulatory change management, contributing to reducing the time and effort needed to update the regulations and make the process more efficient.

Regulation complexity and understanding it are a crucial problem, as evident in the Global Financial Crisis in 2008 [1]. RegTech has since been introduced. However, RegTech alone is insufficient to ensure that regulations are followed and implemented correctly. Moreover, to ensure that regulations are applied in the respective industries with full implementation, it is important to identify all regulations and legislation in the entities that are updated to match the regulatory bodies.

While LLMs are being used in several different industries, they hold an important position in making a great impact on regulatory change management, contributing to reducing the time it takes to update policies and identifying outdated regulations to ensure compliance with current standards and regulations for organizations. As mentioned above, firms must update their policies every year, with a significant amount of time invested in this process. An AI model to complete the same task will help immensely, not only in highlighting outdated sections of the regulations but also in reducing the time and effort it takes to reflect the updates.

It is important to mention that, as the main sources, IEEE, ACM, ScienceDirect, and SpringerLink were chosen. When looking for primary studies, backwards snowballing was used

to find and scan the references of articles that were relevant to the research.

With RegTech and the emerging implications of machine learning, such as generative AI and Machine-Readable Regulations, in the world of legislation and legality, this study focuses on a thorough analysis of Machine-Readable Regulations and their application to machine learning as well as natural language processing, with a focus on generative AI.

This study is structured as follows: Section II describes the objective and contribution of the study in this field of regulatory change management and the significance and potential of ML models, followed by Section III, which focuses on the background study of regulations and compliance and discusses the challenges in interpreting them. It also presents different techniques that have been implemented to convert complex regulations into machine-readable code. Section IV describes in detail the current implications of machine-readable codes of regulations and their benefits and challenges industry-wise. Different techniques for extracting Machine-Readable Regulations and applications of machine-readable codes of regulations are presented in Section V and Section VI. Finally, Section VII presents the conclusion of the survey analysis, and Section VIII discusses the challenges and future directions of machine-readable codes of regulations, offering a guide to researchers.

II. Objectives and Contribution

Regulations are significant business processes that every business and individual must follow to be compliant with regulatory bodies. Regulations and legislative laws can be difficult to understand, although they are written in human language. Not every individual can understand the legal terms and conditions and, therefore, might face difficulty in following them. Again, businesses must be compliant with regulatory bodies to ensure firms are following the changes and updates required to protect rights and the environment and be accountable for their actions. While it takes a significant amount of time to understand the regulations, find duplicates or changes in regulations and update internal policies accordingly, firms and businesses hire many compliance and legal experts to manage these mountains of regulatory data. Therefore, it becomes crucial to automate this time-consuming task. Converting the regulations and legislative laws written in natural language into MRC to identify hidden patterns and duplicates to highlight conflicting regulations can accelerate digitization. ML's influence can play a significant role in reshaping risk and compliance management. An AI model based on fine-tuned LLMs that can highlight and identify hidden patterns and duplicates in conflicting regulations to underline changes and allow businesses and individuals to be up to date with policies can be valuable implementation in the regulatory industry.

This study aims to identify and study existing and emerging ML models to achieve the following objectives:

- Anticipate studying the hidden patterns between the regulatory changes and the outdated policies where changes must be applied.
- Dig deep into the historical patterns to forecast/predict future changes in regulatory compliance with the use of ML models.
- Study the implementation of machine learning to correctly identify the prediction of regulatory changes and correlate them with the actual changes applied by the regulatory bodies.
- Investigate the correlation between the regulatory changes and the impact on the businesses and individuals who must be compliant with the regulations and legislative laws.

III. Background

A. Regulations and Compliance

Regulations are significant business processes that every business and individual must follow to comply with regulatory bodies. Regulations and legislative laws can be difficult to understand, although they are written in human language. Not every individual can understand the legal terms and conditions and, therefore, might face difficulty in following them. Again, businesses must be compliant with regulatory bodies to ensure that firms are following the changes and updates required to protect rights and the environment, and be accountable for the actions that they perform. While it takes a significant amount of time to understand the regulations, find duplicates or changes in regulations, and update internal policies accordingly, firms and businesses hire many compliance and legal experts to manage these mountains of regulatory data. Therefore, it becomes crucial to automate this time-consuming task. Converting regulations and legislative laws written in natural language into MRC to identify hidden patterns and duplicates to highlight conflicting regulations can accelerate digitization. ML influence can play a significant role in reshaping risk and compliance management.

Aligning with national or international regulatory bodies in respective jurisdictions or industries is regulatory compliance. These regulations are internal and involve processes and procedures aimed at streamlining internal business requirements. These regulations not only help in internal processes but also ensure adherence to laws, regulations, guidelines, and specifications relevant to the respective industries. Some examples of regulatory compliance include the Payment Card Industry Data Security Standard (PCI DSS), the Health Insurance Portability and Accountability Act (HIPAA), the Federal Information Security Management Act (FISMA), the Sarbanes-Oxley Act (SOX), the EU's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). As the number of rules increases and the techniques for breaching them increase, many new laws and regulations are developed and introduced so that violations and fraud can be lessened. That said, firms or businesses must also update their old and new regulations and policies when a new law or regulation is passed. This is a manual process that requires the regulatory change management process to be streamlined. Regulatory change management ensures that firms comply with the changes made by the regulatory bodies in the regulations. This activity includes keeping up with the regulatory content,

understanding all the regulations that apply to them, finding out what policies to update and how, etc. This is a complex and time-consuming task for which financial technology (fintech) and financial institutions (FIs) hire many compliance/legal experts to manage mountains of regulatory data.

Regulatory Technology (RegTech) is a technological solution for regulatory compliance and regulations. The main objective of RegTech is to improve regulatory processes. It helps with preparing reports on new and old regulations. One of the objectives of RegTech is to reduce the amount of work completed for manual reporting and to digitize the process. In addition, RegTech helps with risk assessment, identity management and control, transaction monitoring, and fraud prevention, such as money laundering. Furthermore, they address risk assessment [2], [4]–[6]. These are some of the activities that are conducted to ensure regulatory compliance. RegTech aims not to change the industry in which it is implemented but rather to ease the process of regulatory compliance by helping large institutions with the regulatory burden, compliance, and sudden or enormous changes in regulations, and keeping the policies up to date. Regulations usually operate under constant uncertainty, and it becomes more difficult to identify what should be updated to ensure regulatory compliance. This is where RegTech plays a role. FinTech and FIs have started using RegTech to allow themselves to comply with the regulations of regulatory bodies. However, to accelerate digitization, human involvement and interference in RegTech must be minimized, especially in regard to managing regulatory changes.

B. Challenges in Interpreting and Complying with Regulations

As mentioned above, a significant problem lies in regard to regulation complexity and accurately understanding it. As evident in the Global Financial Crisis in 2008 [1], the complexity of regulations played a vital role in the crisis. RegTech has since been introduced. However, RegTech alone is insufficient to ensure that regulations are followed and implemented correctly. Moreover, to ensure that regulations are applied in the respective industries with full implementation, it is important to identify all regulations and legislation in the entities that are updated to match the regulatory bodies.

Along with understanding the regulations, identifying the changes being made to the regulations so that the internal policies can be updated is highly important. This part of managing regulatory changes allows firms to comply with the changes made by regulatory bodies. There are several activities related to this, including understanding the regulatory content, identifying the changes, and determining what internal policies must be updated. For a human, this is a complex and time-consuming task. Firms usually hire many compliance and legal experts to manage mountains of regulatory data, which also becomes an expensive cost.

The number of regulatory changes taking place in organizations is surprisingly large. According to a case study published by iBM [2], the U.S. Code of Federal Regulations (CFRs) contains more than 174,000 pages. It is as long as 25 miles long if printed. The amount of regulatory change has tripled in the last five years, the case study from iBM claims, which was conducted in 2021. This can be a resource-intensive task to understand the complex regulations and take a significant amount of time to reflect on the updates. As a rule, policies should be updated every 13 years [3], [7]. Moreover, high-risk industries such as healthcare, public safety, and financial services are advised to update their regulations more frequently [7], [8]. The time required to update policies depends on multiple factors, including the complexity of the regulations, the scope of the changes needed, the size of the organization, and the number of stakeholders involved. Regular review and updates of regulations are essential to ensure compliance, effectiveness, and alignment with the evolving needs of the business and the external environment.

The potential risks of not updating outdated regulations are significant and can impact various aspects of an organization. It can lead to noncompliance with current standards and regulations [9], [10], security vulnerabilities [11], increased operational costs [9], a lack of clarity and consistency [11], obsolete technology and manual workarounds.

Large Language Models (LLMs) can be used to identify similarities between outdated regulations and highlight the changes needed to match new regulations published by regulatory bodies. This process involves comparing the language and content of old and new regulations and understanding the implications of these changes for various stakeholders. LLMs can play a crucial role in identifying similarities between outdated regulations and highlighting the changes needed to match new regulations published by regulatory bodies. This can help ensure that regulatory compliance processes remain accurate and efficient in the face of rapid technological advancements and changes in regulatory landscapes.

While LLMs are used in several different industries, they hold a prominent position in making a great impact on regulatory change management, contributing to reducing the time it takes to update policies and identifying outdated regulations to ensure compliance with current standards and regulations for organizations. As mentioned above, firms must update their policies every year, with a significant amount of time invested in this process. An AI tool to complete the same task will help immensely, not only in highlighting outdated sections of the regulations but also in reducing the time and effort it takes to reflect the updates.

C. Machine-Readable Data

While machine learning has shown numerous applications in several different industries, RegTech is no different. Businesses and individuals have started using RegTech to allow themselves to comply with the regulations of regulatory bodies. However, to accelerate digitization, human involvement and interference in RegTech must be minimized, especially in regard to managing regulatory changes. It is extremely important to comply with these regulations to speed up the unprecedented customer traffic channeled by businesses. To ensure due diligence, review of customers/businesses (KYC/KYB) and be on top of AML

regulations, businesses and individuals, especially fintech and FIs, are required to conduct periodic checks to establish preventive measures against cyber fraud, money laundering, or terrorist financing activities. This is a complex and time-consuming task for which businesses hire many compliance/legal experts to manage mountains of regulatory data. The influence of ML can play a significant role in reshaping risk and compliance management to accelerate digitization. In addition, regulations can be challenging to understand. Converting them into MRCs and allowing LLMs to be trained and finetuned to achieve a specific specialization can ease the task of complying with regulations. The significant changes in the regulations must be reflected in internal policies for businesses and individuals to follow. It is necessary to ensure consistent governance with regulations.

1) Machine Readable Regulations (MRR): The ability of machines to parse legal documents is what Machine-Readable Regulations are capable of doing. MRR allows access to legal documents and regulations in a machine-readable format. Usually, regulations and policies are drafted and published in digital format, which is easily accessible for humans but not for machines. Machines cannot access legal documents published in digital format. Therefore, it is essential to structure them in a correct form that is accessible to machines, such as XML [1]. In addition, machine-accessible legal documents predefine special delimiters identifying different parts of the document. The term that can be used here is Machine Readable Format (MRF) for documents accessible in this structure, which XML offers. XML can be used to structure any content, not only legal documents. Other examples include the United States Legislative Markup (USLM) used in the US only and the UN's Akoma Ntoso MRF, which can be considered an international MRF [1].

2) Semantic Machine-Readable Regulations (SMRR): MRR focuses on the structure and accessibility of legal documents, which makes the content readable to machines. However, it is important to understand the meaning of the content that Semantic Machine-Readable Regulations (SMRR) allows. The main focus of SMRR is to have semantics for the structured content to make it useful and semantically machine-readable. This semantics produces internal and external insights. Internal insight allows us to understand various possible meanings of the content itself. However, external insight explains the relationship between the provisions of the document and its corpus. These insights provide metadata that machines can access. X2RL and JSON formats are examples of SMRR formats [1].

3) Machine-Executable Regulations (MER) and Machine-Consumable Regulations (MCR): While ensuring that machines have access to regulations is important, it is also important to carry out functions using these metadata. Functions include generating compliance reports, monitoring, and updating regulations. Codifying legal documents is one of the means of achieving the function offered by Machine-Executable Regulations (MER) and Machine-Consumable Regulations (MCR). The codification allows the underlying meaning of the regulations to be converted into a code that machines can execute. Regulations or legislation can be Boolean logic or conditional statements that can be executed with one or several lines of code. MERs and MCRs are extensions of SMRRs. Unlike SMRRs, MERs and MCRs do not produce metadata, but they execute MCR content [1].

4) Rules as Code (RaC)/Legislation as Code (LaC): MRR, SMRR, and MER/MCR are concepts and traits of Rules as Code (RaC) [equally Legislation as Code (LaC)]. The main difference between the previously explained concepts and RaC is integrability. Breaking down the legal corpus into basic legal items and encoding them into basic modules are the first steps of RaC. These modules then accept structured input and produce structured output. The RaC document integrates the set of modules in a constructed document, ensuring that there are no repetitions of legal concepts that can be possible in MER/MCR, since they codify legal concepts only without considering coding repetition, such that some legal items that are common across documents are coded several times separately [1].

D. Generative AI (GenAI)

GenAI is a type of AI that can create new content, such as text, images, and music. It can be used to generate realistic synthetic data, which can be used to train other AI models or to generate new products and services. In the context of regulatory compliance, GenAI could be used to generate realistic synthetic datasets of regulations, which could be used to train AI models to identify conflicting regulations and gaps in regulations. On the basis of several types of inputs, Generative AI (GenAI) creates new content. This content can include text, images, sounds, animations, 3D models, etc. GenAI uses Neural Networks (NN) to identify hidden patterns and structures within any type of data and generate new content. The training of GenAI is not bound to any one of the unsupervised or semi-supervised learning approaches. However, it can be learned and trained in any way. Therefore, organizations can leverage GenAI's features, as it allows them to create foundation models because of the large amounts of unlabeled data. Given that foundation models can be versatile, they are AI-based systems or services whose usage is decided in which industry it is implemented. Some examples of GenAI include GPT-3 and Stable Diffusion [12].

There are five algorithms that provide the framework to develop the GenAI system [13].

1) Overview of generative AI techniques

a) Autoencoders: Autoencoders are a type of neural network that does not require training with prelabeled data [14]. The main use of autoencoders is to compress high-dimensional data into lower-dimensional data. This is called a latent space. The encoder and encoder components are jointly trained. However, because of the low feasibility of AE designs, the change in the latent variables is not high. This is where the Variational Autoencoder (VAE) is introduced, allowing more constraints to be added to latent variables for the desired distribution. While VAEs recognize problems with low feasibility, they produce blurry outputs [13].

b) Generative Adversarial Networks (GAN): A GAN is a deep learning architecture that trains two neural networks to produce authentic data from a given training dataset [15]. The two different networks compete with each other. One network modifies the input data sample as much as possible to produce new data. The other network predicts whether the generated new data belongs to the original input data. The system continues generating new and improved versions of data until it is no longer related to the original input data.

c) Autoregressive model: With the exceptional capability of density estimation, Autoregressive Models sequentially predict each variable component on the basis of prior elements [16]. The performance is affected by the slower processing of a high volume of data. One of the advancements of autoregressive models is the integration of self-attention mechanisms, especially transformer models. This allows parallel processing and selective focus on sequence elements.

d) Diffusion and flow-based model: Differences in operational mechanism, diffusion, and flow-based models are quite similar and often grouped. Diffusion models add noise to data and conduct reverse engineering for data reconstruction [16]. It consists of forward and backward diffusion processes where forward noise is introduced and backward diffusion diminishes this noise. Furthermore, flow-based models leverage normalizing flow to generate a probability distribution between the data and the latent space. This is particularly useful for density estimation.

e) Foundation model (FM): FMs are foundational models that are trained on vast amounts of data and can produce highly accurate and diverse outputs [16]. This can range from images to texts to complex simulations. They can generate very specific tasks on the basis of limited datasets for improved performance. Some examples include image and text generation, multimodal models, and reinforcement learning models.

IV. Machine-Readable Code of Regulations (MCR)

Machine-readable code (MCR) has become a booming field of research in which the focus is mainly centered on how any text can be converted in such a way that it can be easily understood, processed, and parsed by a computer. In regard to MCRs, one area that has gained the attention of regulators is the use of MCRs in regulations. It becomes quite difficult for industries such as FinTech, Financial Institutes, Healthcare, etc., to keep up with the updates and changes in regulations by regulatory bodies and ensure that they are compliant with all the policies and procedures. Machine learning (ML) can help computers process these regulations and make this tedious task easier for the respective compliance departments of different industries.

Machine learning (ML) can add significant value to the regulatory industry in regard to abiding by regulations and ensuring that other businesses and individuals are compliant with existing and new regulations. For example, programs using ML algorithms are already being implemented in the field to detect fraud and anti-money laundering activities or monitor transactions in the financial industry [17]. The ML is trained such that it can read Machine Readable Code (MRC), which has been encoded from the natural language [17]. Furthermore, converting regulations written in natural language is one of the main activities allowing an ML to be trained on this data to support regulatory compliance. Researchers have suggested and offered different approaches to convert natural language into MRC. Some of the prominent forms of machine-readable formats for regulatory documents are XML, JSON and X2RL. The process of converting to a machine-readable format is important to extract the correct meaning of the regulations, which might sometimes follow inconsistent language. In [18], the authors introduced a new machine-readable format and developed X2RL to enhance the semantic representation of regulatory information. The study shows that X2RL helps in enabling a more granular and precise representation of regulatory documents and capturing the underlying semantics. It also discusses how semantic representation can aid in automated interpretation and analysis of regulations, which is a requirement for the use of ML algorithms for regulations. Moreover, [19] introduces a "legislative recipe" as the idea of a structured syntax designed to represent legislative texts in a machine-readable format. This design includes provisions, definitions, and references that enable precise and standardized representation of the legislative components. It also accommodates the semantic annotation and hierarchical nature of the regulations. In a previous paper [20], the authors described the foundational concept of the JSON schema and how it can be a tool for describing and validating the structure of JSON data. It helps specify the structure, constraints, and validation rules for JSON documents, which leads to the validation of common types of strings, numbers, and booleans. In [12], the authors aims to improve the encoding of copyright legislation into machine-executable code. To ensure accuracy in "Rules as Code", it is important to involve legally trained individuals. This study revealed that collaborative agreement on key legal terms leads to better-encoded rules. However, different interpretive choices can still occur due to complexities in statutory interpretation and coding language functionality. The authors suggest separating technical validation and legal alignment processes to improve accuracy. The authors used Defeasible Deontic Logic (DDL) and Turnip to convert selected provisions of the Copyright Act into machine-executable code. Furthermore, [21] discussed the inherent complexity of regulatory texts, which are often filled with legal jargon, ambiguities, and references to other regulations. They emphasize the challenges this complexity poses for both human understanding and machine automation. The authors delve into how natural language processing (NLP) and machine learning techniques can be employed to interpret regulatory texts. The authors further emphasized the point that rulemaking allows the automation of compliance, reporting, and enforcement processes with the help of ML technologies.

The research field is booming with several different types of approaches to achieve the goal of an efficient and faster process of converting regulations into MRCs. While it is challenging to convert natural language into machine-readable code, several machine-learning algorithms or techniques have been developed to aid in MRC. The task of converting into MRC falls under the domain of Natural Language Processing

(NLP) and Natural Language Understanding (NLU). Several machine learning algorithms and techniques can be employed for this purpose, including Recurrent Neural Networks (RNNs) [22], Transformer Models [23], Seq2Seq Models [24], Graph Neural Networks (GNNs) [25], and Large Language Models (LLMs) [26]. The authors of [22] reported that an RNN can be used to recognize written text and convert it into MRC. They used the fundamentals of Long Short Term Memory (LSTM) to capture the long-range sequence of handwritten elements to identify noise in handwritten data and segments, and isolate it into individual characters. This study also presents the implementation of Seq2Seq models where the input is the sequence of handwritten texts and produces corresponding text sequences, but with MRC as the output. Moreover, [23] highlighted the importance of an ML technique in understanding software code and generating related and useful comments for specific code snippets. If an ML technique can convert an MRC into natural language, it is an opportunity for researchers to leverage the benefits of these MLs and produce state-of-the-art results in natural language understanding for the MLs. The main goal and result of the study [23] was to use semantic parsing in the context of programming languages and extract meaningful information from code. The authors presented the "Setransformer" architecture, which was specifically designed for the semantic parsing of code. Leveraging transformer models to capture both the structural and semantic information within the code helps in generating code comments in human-like language. The authors of the paper [24] presented a thorough analysis of how the GNN can be leveraged for natural language processing. Exploring different ways of representing natural language, such as graphs, dependency trees, co-occurrence graphs, and knowledge graphs, this study highlights the advantages of using graph-based representations for capturing semantic relationships in language. The survey covered a wide range of NLP tasks where the GNN has been applied, such as sentiment analysis, named entity recognition (NER), relation extraction, question answering, knowledge graph completion, language modelling, and semantic role labelling.

In practice, recent studies have discussed the significance of machine learning (ML) in regulatory change management, particularly in the context of converting regulations into Machine Readable Code (MRC). The implementation of ML algorithms in regulatory industries, such as fraud detection and anti-money laundering, is highlighted. ML models are trained to read MRCs encoded from natural language regulations. There are various approaches for machine-readable formats, including XML, JSON, and X2RL. The complexity of regulatory texts, filled with legal jargon and ambiguities, is acknowledged, and the use of natural language processing (NLP) and machine learning techniques for interpretation is discussed.

A. Objectives and Benefits of Adopting MRCs

1) Efficiency Improvement: The automated extraction of key information from regulatory documents saves time and effort for the compliance department, which is engaged in human interpretation, to identify the meaning of the information, diving deep into the regulations and applying them to the workplace. This makes the compliance process more efficient. As mentioned in [27], a significant amount of time is spent on understanding and identifying areas of change in regulatory policies. Given that MRC can read and process regulations without the involvement of a human resource, it not only saves time but also makes it easier for humans to interpret regulations and understand complex clauses, which must be updated every one to three years [3], [7].

2) Accuracy enhancement: Legally trained individuals are required to ensure the accuracy of policies and regulatory documents. Additionally, each individual interprets the policies in different interpretive choices, which creates statutory interpretation and coding language functionality. The authors [21] suggested that NLP can increase accuracy and eliminate the complexity of understanding regulations by converting them to MRCs. The authors further emphasized the point that rulemaking allows the automation of compliance, reporting, and enforcement processes with the help of ML technologies. The MRC helps minimize the risk of human error, ensuring that compliance activities are executed accurately and in strict accordance with regulatory requirements.

3) Real-time monitoring and adherence: MRC enables monitoring of regulatory changes and allows identification of the areas where the changes have been reflected. The objective of this ability of MRC is to create consistency between the policies implemented in the respective fields and ensure adherence to the regulatory updates by the regulatory bodies. This allows us to adapt to modifications to regulations and maintain adherence.

4) Standardization of compliance processes: A uniform approach is required to be compliant with various jurisdictions and sectors in which the respective compliance department individuals invest their resources, effort, and time. MRC promotes the use of standardized formats and structures for regulations, making it easier to create and understand complex regulations.

5) Automation of reporting: With MRC, an automated process of reporting can be introduced by offering a standardized compliance process, identifying accurate data, and interpreting regulations via MRC. Reporting requires a significant amount of data analysis to abide by regulations accurately. The MRC can ensure adherence to regulations and policies reflecting compliance reporting.

B. Evolution of Machine-Readable Regulations (MRRs) and Current Development [1]

Access to data, meaning that converting data in a standardized form that computers can process, is one of the cornerstones and requirements of MCRs. Computers with the ability to parse legislative content and regulations are quite challenging. Usually, regulations are written in electronic formats, such as MS Word and PDF, which are easy for the human eye to read but not easy for machines to process and identify the meaning of the regulations. To address the challenges of machines being able to access legal content, the

MRR was formed. The readability of regulations for machines is affected by the structure of the regulations or the legal contents. This means that the format of the content must be in a way that helps to discern the components so that the identification of the components is similar to how a human does. This structure is referred to as Machine Readable Format (MRF), and one example of such a format is XML. It can be used to structure any content, not only legalese or regulations. MRF is the first of its kind that helps in transforming legal content into a form that can be processed and parsed by machines. This is an essential step toward RaC. While MRR focuses on the structure of regulatory documents, SMRR focuses on the enrichment of texts as well. The process of SMRR includes the injection of semantics in the machine-readable structure in the form of metadata. This is similar to how a human interprets a document and drafts the semantics. To further improve the process, codification has been added to the process of parsing regulations by machines. MER and MCR allow converting the documents into Boolean logic and further converting it to computer code, which is called codification. To conduct the process of codification, the text must be structured and enriched, which is also part of the MER and MCR. To ensure the integrability of the documents, the document is broken into a corpus that consists of basic rules and modules. These different modules are then converted into machine-readable code, ensuring that the same meaning of different modules is not repeated during the parsing process. This process is called RaC and LaC.

C. Common Standards and Machine-Readable Formats

1) JavaScript Object Notation (JSON): The JSON schema provides a standardized schema language for JSON documents [20]. It allows developers to specify the structure of JSON documents, constrain their content, and verify the integrity of requests and responses in APIs. The JSON schema is used for various purposes, including data validation, document-oriented databases, automated document generation, and schema validation tools. JSON is commonly used as a machine-readable format since it is lightweight and easy to parse for machines. It is ideal for transmitting data between systems.

2) XML: XML is one of the most prominent machine-readable formats [1], [18]. The US's regulations, named the United States Legislative Markup (USLM), and the internationally recognized regulation, called the UN's Akoma Ntoso, are XML-based. These XML-based regulations contain a hierarchy of tags and attributes. It helps improve the internal structure of the legal document rather than the content. They sufficiently lack the developed methods of connecting the bodies of legal documents in a machine-readable way. This helps reduce the management cost of legal documents; however, the main goal is to minimize the economic cost of legislation for those who must comply with the body of the regulations.

3) eXtensible Regulatory Reporting Language (X2RL): Authors in [18] proposed a new machine-readable format that, along with having tags and attributes such as XML, has additional new attributes, tags, and models designed to enrich the metadata of the legal document's content. According to the authors, X2RL helps in processing the legal content, intent, scope, and meaning of the legislation and regulations. In addition, it identifies the depth of the external structure of the documents to link multiple documents together, which reduces the economic cost and allows efficient data interchange. The authors introduced <**provision**> as the basic container for any document fragment. Strengthening the external structure of the legislation and expanding the ontology to include new elements to be able to describe the legal contents in detail are two more goals X2RL achieves.

4) YAML Ain't Markup Language (YAML): YAML is also a markup language used to provide a structure for data [28]. Like JSON, YAML stores data in a human-readable format by providing serialization and offering text marking. YAML can be used in place of JSON since it is now a subset of JSON. Supporting scalar datatypes such as lists and arrays, it stores data in a text file.

5) Resource Description Framework (RDF): RDF is mainly used to convert data on the web in a machine-readable format [29]. In RDF, information is represented in the form of triples, which consist of subject-predicate-object statements. The subject and object represent resources, whereas the predicate indicates the relationship between them. Facilitating data integration and interoperability, RDF is commonly used to create linked data.

6) Web Ontology Language (OWL): Like RDF, OWL is used for data on the web. It represents domain knowledge via classes, properties, axioms, and instances for use in a distributed environment [30]. Representing a semantic representation of domain knowledge, OWL supports efficient reasoning and expressive power. The authors of [31] proposed a new method of generating OWL ontologies from XML data sources to highlight how XML data can be extracted to represent domain knowledge in OWL format.

D. MRR Use Cases

Converting legal content and regulations into machine-readable language has made strides in the RegTech industry. It is a transformative movement for the global regulatory community to ease the complexity of regulations, legislation, and policies, allowing entities to assist in complying with new regulations and drafting or updating old legislation and policies. RegTech can be further improved with the essence of MRR so that machines can analyze legal content to parse the content itself. The following are some of the use cases that are framed by MRRs and offer improvements in the regulatory industry.

1) Committed to regulations: MRR adaptation is effective and can be an efficient way of ensuring that an entity is committed to following regulations. Regulators can easily identify if entities are misinterpreting regulations and making erroneous decisions. MRR allows entities to refrain from making any judgment errors. It improves the quality of compliance. An MRR implementation helps in the use of machine learning algorithms to study the regulations and

determine whether they are correctly implemented and followed within the organization.

2) Regulations updates: MRR does not merely translate the regulations into machine-readable language but also creates the opportunity to use machine learning algorithms to identify the updates in regulatory bodies and how existing regulatory policies must be updated. It helps highlight outdated rules and update them with new implementations. Financial institutes can prepare MRR-based handbooks and search places where humans and machines can search for specific rules and regulations. MRR can also help in identifying duplicate policies within an organization and enhancing them by removing duplicates and replacing them with the correct format. This signifies and symbolizes the movement of machine-analyzable legal content and the process of generating legal documents.

3) Implementation of principle-based regulations: To facilitate measurements of the competency of licensing requirements for financial institutes and ensure that they meet supervisory expectations, following regulations is important. MRR allows the implementation of regulations to meet the requirements of the regulatory body by aggregating relevant documents for analysis and comparing them with what is being followed within the organization/entity. A principle-based regulation requires a greater level of understanding by machines, which is needed for the compliance of financial institutions with regulatory bodies. An extensive MRR adaptation in the financial sector to meet licensing requirements is a significant and effective approach.

4) Financial reporting: With the help of MRR, the reporting process can be much more efficient and faster. Using technology, the current reporting process can be automated to be more accurate and consistent by standardizing the description and identification of data, as well as digitizing reporting instructions and generation through MRR. Reporting is crucial in regard to abiding by regulations in the right way. The MRR can ensure 100% adherence to regulations and policies and reflect the same in financial statements and compliance reporting.

V. Techniques for Extracting Machine-Readable Regulations

MRR allows machines to parse complex legislation, regulations, and policies that humans might find difficult to interpret. It not only converts regulations into machine-readable formats but also offers the chance to identify areas of change, ensuring the compliance of policies and rules that a body is following with regulatory bodies. NLP offers extensively utilized techniques to extract and parse regulations from textual documents in various domains, such as legal, compliance, governance, and policymaking. As mentioned earlier, this practice of converting regulations into machine-readable formats involves identifying and understanding rules, interpreting them for machines matching the human meaning, and allowing different sectors to follow the regulations properly. There are multiple techniques to extract text from regulations so that legislation can be considered Machine-Readable Regulations.

1) Text processing: Texts must undergo preprocessing steps such as tokenization [31], stop word removal [32], and stemming or lemmatization to clean and normalize the text. To conduct NLP tasks accurately and efficiently, it is crucial to analyze and process texts.

2) Named Entity Recognition (NER): Once the text is preprocessed, there are multiple ways of extracting regulations via NLP. One of them is Named Entity Recognition (NER) [33]. It involves identifying and categorizing entities mentioned in the text into predefined categories such as organization names, dates, locations, and, most importantly, regulatory entities such as laws, rules, and directives. According to [33], NER focuses on recognizing particular designators from texts. It requires predefined semantics such as person, location, and organization. It not only extracts information but also plays an essential role in applications such as text understanding, information retrieval, automatic text summarization, question answering, machine translation, and knowledge base construction. There are several techniques applied in NER, including rule-based approaches, unsupervised learning approaches, feature-based supervised learning approaches, and deep learning-based approaches.

3) Rule-based approaches: Rule-based approaches do not need any annotated data. They are based on handcrafted rules [34]. The rules are designed on the basis of domain-specific semantics and syntactic–lexical patterns. They capture specific patterns or formats of regulations within the text. Since this approach focuses on a specific domain, knowledge becomes difficult to transfer to a different domain system.

4) Topic modelling: Topic modelling is an analytical tool used to evaluate data [34]. It focuses on the main topics or themes of the textual contents and helps identify them. It becomes easier to understand and extract themes of the topics of the regulatory documents, and make it easier to determine the meaning of each. Some of the topic modelling approaches include Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

5) Text classification: As the name suggests, text classification is a model trained to categorize textual documents into different domains or types. This approach might seem similar to topic modelling, but text classification focuses on the broader area of the regulations, segregating them into regulations for each sector where the regulation is applied [35]. Examples include financial regulations, environmental regulations, and healthcare regulations. Supervised learning algorithms, such as Support Vector Machines (SVMs) or deep learning architectures like Convolutional Neural Networks (CNNs), can be utilized for text classification.

6) Semantic analysis: Another NLP approach for Machine-Readable Regulations is semantic analysis. It helps in understanding the meaning and context of regulatory text [36]. It focuses on the relationships between words and

phrases to interpret and infer the intent and implications of regulations.

7) Information extraction: While it becomes difficult to extract the meaning of complex regulations, information extraction approaches are applied to extract structured information from unstructured regulatory text [37]. The extraction process involves identifying key elements such as obligations, permissions, prohibitions, and conditions mentioned in the regulations. The entire process includes sentence segmentation, tokenization, part-of-speech tagging, entity reorganization, entity disambiguation, relation extraction, and event extraction.

8) Knowledge graph: While the information extraction approach focuses on extracting a structured format of regulations from unstructured regulations, knowledge graphs are constructed to represent the relationship between the structured and interconnected formats of extracted texts [38]. The extracted entities are organized in a graph-like structure that represents the relationships and attributes of the regulations.

TABLE I. TECHNIQUES FOR EXTRACTING MACHINE-READABLE REGULATIONS (MRR)

Technique	Purpose	Methodology	Advantages	Review
Text Processing [31], [32]	Cleans and normalizes text for NLP tasks.	Includes tokenization, stop-word removal, stemming, and lemmatization.	Improves accuracy and efficiency of downstream NLP techniques.	Requires domain adaptation to handle legal language nuances [39].
Named Entity Recognition (NER) [33]	Identifies and classifies key entities in regulatory text (laws, rules, organizations, etc.).	Uses predefined semantic categories (e.g., person, organization, law). Methods include rule-based, unsupervised, supervised, and deep learning approaches.	Enables extraction of relevant regulatory entities for structured representation.	Rely on predefined semantics.
Rule-Based Approaches [34]	Extracts regulatory patterns using handcrafted rules and domain-specific semantics.	Relies on syntactic and lexical patterns defined by experts.	Effective for well-defined, narrow domains with consistent language.	Approach focuses on a specific domain; knowledge becomes difficult to transfer to a different domain system.
Topic Modelling [34]	Identifies hidden themes and major topics in regulatory documents.	Uses algorithms such as LDA and NMF.	Helps understand and categorize large volumes of regulatory text by themes.	May not capture context or subtle legal relationships; requires post-interpretation [40].
Text Classification [35]	Categorizes regulations into domains (e.g., financial, environmental, healthcare).	Employ supervised learning models such as SVM or deep learning (CNNs).	Enables domain-specific regulation management and retrieval.	Requires labeled datasets as they categorized by sector
Semantic Analysis [36]	Interprets the meaning and intent of regulations.	Analyzes relationships between words and phrases to infer implications.	Provides deeper understanding of regulatory intent; improves contextual interpretation.	Complex to implement; dependent on advanced linguistic and contextual models [41].
Information Extraction (IE) [37]	Convert unstructured regulatory text into structured data.	Involves segmentation, tagging, entity recognition, disambiguation, relation, and event extraction.	Identifies obligations, permissions, and conditions for compliance automation.	Challenging for long, nested legal clauses requires hybrid rule-based and ML methods [42].
Knowledge Graphs [38]	Represents structured relationships among extracted regulatory entities.	Organizes entities and relationships into interconnected graph structures.	Enables reasoning, querying, and visualization of regulatory dependencies.	Building and maintaining accurate graphs requires ongoing data validation and ontology design [43].

Table I presents a summary of the different techniques for extracting MRRs. Mainly, these techniques form a layered approach to converting unstructured regulatory texts into structured, machine-readable formats. The several ways of extracting MRRs are to prepare and identify key elements of the text, extract patterns, categorize and organize the data, and lastly, understand meaning, relationships, and build structured representations. They can be grouped as foundational, such as text processing and NER, ensuring clean, labelled data for interpretation. Intermediate, including rule-based extraction, topic modelling, and text classification, providing domain organization and thematic clarity. Finally, more advanced approaches are semantic analysis, information extraction, and knowledge graph, extending these efforts into contextual understanding and structured representation. They establish a pathway for transforming unstructured regulatory texts into interpretable, machine-actionable formats that support automation, consistency, and scalability in compliance management.

VI. Applications of Machine-Readable Code of Regulations in Machine Learning

A. Regulatory Compliance Automation

Through NLP, the extracted texts from regulations providing a structured format of data can be analyzed and processed by machine learning algorithms. ML can identify patterns between historical regulatory data and the current application of regulations to ensure that industries are compliant with regulatory bodies. In addition, ML models can extract relevant information from regulatory documents to generate compliance reports automatically, reducing manual human work and effort. ML can dynamically monitor regulatory changes by keeping itself up to date with regulatory bodies and identifying areas of change in the current implementation of regulations in the respective industry.

B. Risk Assessment and Predictive Analytics

As ML has shown a proven track record of generating predictive outcomes, the application of ML in RegTech is no different. With the help of enough data and training, ML can predict the future development of regulations and assess their potential impact on businesses, providing enough time for regulatory reform. It also allows firms to anticipate compliance risks.

C. Legal Research and Policy Analysis

ML algorithms help summarize complex regulations and extract key insights by analyzing regulatory data with the help of NLP. It offers identifying trends and patterns in the legal industry and aids legal research and policy analysis.

D. Industry-Specific Applications

While RegTech is highly used in the financial industry, its limits do not stop there. In addition to other industries, healthcare and manufacturing are two prominent areas where RegTech is being implemented and has shown great results. In addition to ML algorithms offering automated compliance monitoring, risk assessment, and reporting in the financial sector, ML techniques can help ensure that healthcare industries are compliant with their industry-specific requirements, such as HIPAA in the US. Furthermore, ensuring adherence to safety, environmental, and quality standards, which can be ensured by ML given its impact on RegTech in all aspects of ensuring regulatory compliance, is highly important for the manufacturing industry.

VII. Conclusion

The growing complexity of regulatory frameworks has made compliance both essential and intensive. This study highlights how translating regulatory texts into Machine-Readable Regulations (MRR) can transform compliance from a largely manual, interpretive task into a structured, data-driven process. By enabling machines to interpret and act on regulatory information, MRR offers clear advantages like greater efficiency, consistency, and adaptability in managing compliance across sectors. Beyond automation, integrating Machine-Readable Code (MRC) with AI and knowledge-based systems opens new possibilities for regulation management. Tools like GenAI, ontologies, and federated learning illustrate how compliance can evolve toward predictive, privacy-conscious, and continuously improving systems. Finally, the key insight is that regulatory compliance can no longer rely solely on human interpretation or static systems. The future lies in collaboration between technology and regulatory institutions, where agile frameworks and machine-readable models allow compliance to evolve in step with technological and policy change.

VIII. Challenges and Future Directions

A. Data Quality and Standardization

To ensure that the ML is trained well and that the MRR can generate the expected outcomes, it is important to train the algorithm on complete and consistent data. In addition, a standardized format of training to ensure regulations across different jurisdictions and domains allows the MRR to enhance the quality of its results, promoting the interoperability of data. Advances in semantic technologies, such as ontologies and knowledge graphs, can facilitate the semantic interoperability of Machine-Readable Regulations, enabling more precise and context-aware interpretation.

B. Privacy and Ethical Considerations

The MRR may contain sensitive and confidential information, hindering the privacy of the data. In addition, MRR may exhibit biases leading to unfair outcomes in regulatory compliance decisions on the basis of the training dataset provided to the ML algorithm. Given that the implications of MRR are not limited to a specific industry, it may reveal information affecting the governance framework. Therefore, techniques must be implemented to preserve the privacy of data by utilizing differential privacy and federated learning in MRR. Furthermore, ML algorithms can be developed to indicate bias detection to ensure that fair and ethical guidelines can be introduced in industries and regulatory bodies to preserve the governance framework.

C. Emerging Technologies and Trends

Interdisciplinary collaboration is required so that MRRs can keep up with the emerging changes in the world of technology and ensure that novel regulatory approaches are captured by them. Moreover, regulatory bodies can adopt agile and adaptive regulatory frameworks leveraging MRRs to enable rapid responses to technological advancements and regulatory changes.

Acknowledgment

We acknowledge that a preprint version of this manuscript is available on a non-commercial and academic site.

References

[1] Communications, Space & Technology Commission (CST), "Machine Readable Regulations and Beyond: Tackling Regulatory Complexity," 2023. [Online]. Available: https://www.cst.gov.sa/ar/mediacenter/researchsandstudies/Documents/Tackling_Regulatory_Complexity_Machine_Readable_Regulations_and_Beyond.pdf

[2] Deloitte, "Inside CCO, CISO, CRO, CIA, BOD: Insights from Deloitte," 2017. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/about-deloitte/Inside/lu_inside-global-edition-2017-full.pdf

[3] ComplianceBridge Team, "Have You Been Regularly Updating Policies and Procedures?" ComplianceBridge, November 8, 2022. [Online]. Available: https://compliancebridge.com/updating-policies-and-procedures/

[4] D. W. Arner, J. Barberis, and R. P. Buckey, "FinTech, RegTech, and the reconceptualization of financial regulation," Northwestern Journal of International Law & Business, vol. 37, p. 371, 2016.

[5] D. W. Arner, J. Barberis, and R. P. Buckley, "FinTech and RegTech in a Nutshell, and the Future in a Sandbox," CFA Institute Research Foundation, 2017.

[6] D. Yang and M. Li, "Evolutionary approaches and the construction of technology-driven regulations," Emerging Markets Finance and Trade, vol. 54, no. 14, pp. 3256-3271, 2018.

[7] Elliott Davis, "Is it time to review your Policies and Procedures Manual?" Elliott Davis, April 18, 2022. [Online]. Available: https://www.elliottdavis.com/is-it-time-to-review-your-policies-and-procedures-manual/

[8] PowerDMS, "Why it is important to review policies and procedures," PowerDMS Policy Learning Center, December 22, 2020. [Online]. Available: https://www.powerdms.com/policy-learning-center/why-it-is-important-to-review-policies-and-procedures

[9] A. Nadkarni, "How Often Should You Review Your Policies and Procedures?" 24by7Security, 2019. [Online]. Available: https://blog.24by7security.com/how-often-should-you-review-your-policies-and-procedures

[10] C. Anderson, "What are the Risks of Not Updating Policies and Procedures?" Bizmanualz, May 22, 2023. [Online]. Available: https://www.bizmanualz.com/writing-policies-and-procedures/what-are-the-risks-of-not-updating-policies-and-procedures.html

[11] M. Prijic, "Risks of Using Outdated Operating System," IT Convergence, April 6, 2023. [Online]. Available: https://www.itconvergence.com/blog/risks-of-using-outdated-operating-system/

[12] A. Witt, A. Huggins, G. Governatori, and J. Buckley, "Converting copyright legislation into machine-executable code: Interpretation, coding validation and legal alignment," in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL), 2021, pp. 139-148.

[13] T. Sakirin and S. Kusuma, "A Survey of Generative Artificial Intelligence Techniques," Babylonian Journal of Artificial Intelligence, 2023, pp. 10-14.

[14] J. T. Chien, Source Separation and Machine Learning. Academic Press, 2018.

[15] M. Salvaris, D. Dean, and W. H. Tok, Deep Learning with Azure: Building and Deploying Artificial Intelligence Solutions on the Microsoft AI Platform. Apress, 2018.

[16] B. Decardi-Nelson, A. S. Alshehri, A. Ajagekar, and F. You, "Generative AI and Process Systems Engineering: The Next Frontier," arXiv preprint arXiv:2402.10977, 2024.

[17] E. Micheler and A. Whaley, "Regulatory technology: replacing law with computer code," European Business Organization Law Review, vol. 21, pp. 349-377, 2020.

[18] P. A. McLaughlin and W. Stover, "Drafting X2RL: A Semantic Regulatory Machine-Readable Format," MIT Computational Law Report, 2021.

[19] M. Ma and B. Wilson, "The legislative recipe: Syntax for machine-readable legislation," Northwestern Journal of Technology & Intellectual Property, vol. 19, p. 107, 2021.

[20] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, "Foundations of JSON schema," in Proceedings of the 25th International Conference on World Wide Web (WWW), 2016, pp. 263-273.

[21] J. Mohun and A. Roberts, "Cracking the code: Rulemaking for humans and machines," 2020.

[22] I. Paul, S. Sasirekha, D. R. Vishnu, and K. Surya, "Recognition of handwritten text using long short-term memory (LSTM) recurrent neural network (RNN)," in AIP Conference Proceedings, vol. 2095, no. 1. AIP Publishing, April 2019.

[23] Z. Li et al., "Setransformer: A transformer-based code semantic parser for code comment generation," IEEE Transactions on Reliability, vol. 72, no. 1, pp. 258-273, 2022.

[24] L. Wu et al., "Graph neural networks for natural language processing: A survey," Foundations and Trends in Machine Learning, vol. 16, no. 2, pp. 119-328, 2023.

[25] L. Wu et al., "Graph Neural Networks for Natural Language Processing: A Survey," vol. 16, no. 2, pp. 119–328, Jan. 2023, doi: https://doi.org/10.1561/2200000096.

[26] Nikitas Karanikolas, E. Manga, Nikoletta Samaridi, E. Tousidou, and M. Vassilakopoulos, "Large Language Models versus Natural Language Understanding and Generation," Nov. 2023, doi: https://doi.org/10.1145/3635059.3635104.

[27] "Regulatory Change Management Efficiency, Effectiveness & Agility in Regulatory Change Processes STRATEGY PERSPECTIVE Governance, Risk Management & Compliance Insight," 2021. [Online]. Available: https://www.ibm.com/downloads/cas/MAN6JNY4

[28] A. Loibl, T. Manoharan, and A. Nagarajah, "Procedure for the transfer of standards into machine-actionability," Journal of Advanced Mechanical Design, Systems, and Manufacturing, vol. 14, no. 2, p. JAMDSM0022, 2020.

[29] N. Haider and F. Hossain, "CSV2RDF: Generating RDF data from CSV file using semantic web technologies," Journal of Theoretical and Applied Information Technology, vol. 96, no. 20, pp. 6889-6902, 2018.

[30] G. Antoniou and F. V. Harmelen, F. V. "Web ontology language: Owl," Handbook on ontologies, pp. 91-110, 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.

[31] S. J. Mielke et al., "Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP," arXiv preprint arXiv:2112.10508, 2021.

[32] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," in Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 466-472, doi: 10.1109/ICACCS48705.2020.9074166.

[33] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, 2020.

[34] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," Information Systems, vol. 94, p. 101582, 2020.

[35] V. Dogra et al., "A complete process of text classification system using state-of-the-art NLP models," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[36] S. A. Salloum, R. Khan, and K. Shaalan, "A survey of semantic analysis approaches," in Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), Springer, 2020, pp. 61-70.

[37] S. Singh, "Natural language processing for information extraction," arXiv preprint arXiv:1807.02383, 2018.

[38] I. Mondal, Y. Hou, and C. Jochim, "End-to-end NLP knowledge graph construction," arXiv preprint arXiv:2106.01167, 2021.

[39] D. P. Karanikolas et al., "Strengths and Weaknesses of LLM-Based and Rule-Based NLU Systems," Electronics, vol. 14, no. 15, 2025.

[40] K. Ashihara, "Improving topic modeling through homophily for legal documents," Applied Network Science, vol. 5, 2020.

[41] R. M. Bakker, A. J. Schoevers, van Drie, M. P. Schraagen, and T. de, "Semantic role extraction in law texts: a comparative analysis of language models for legal information extraction," Artificial Intelligence and Law, Mar. 2025, doi: https://doi.org/10.1007/s10506-025-09437-x.

[42] D. Premasiri, T. Ranasinghe, R. Mitkov, et al., "Survey on legal information extraction: current status and open challenges," Knowledge & Information Systems, vol. 67, no. 12, pp. 1–35, 2025.

[43] C. Peng et al., "Knowledge Graphs: Opportunities and Challenges," Artificial Intelligence Review, 2023.