frontiers | Frontiers in Built Environment

TYPE Methods
PUBLISHED 28 April 2025
DOI 10.3389/fbuil.2025.1575913

OPEN ACCESS

EDITED BY
Salman Azhar,
Auburn University, United States

REVIEWED BY
Rehan Masood,
Otago Polytechnic, New Zealand
Roy Lan,
University of Texas at San Antonio, United States

*CORRESPONDENCE
Nazym Shogelova,
nazymshogelova@gmail.com

RECEIVED 13 February 2025
ACCEPTED 16 April 2025
PUBLISHED 28 April 2025
CORRECTED 09 June 2025

CITATION
Kabzhan Z, Shakhnovich A, Gorshkov S, Yemenov Y, Gorshkov F and Shogelova N (2025) Semantic and ontology-based analysis of regulatory documents for construction industry digitalization.
Front. Built Environ. 11:1575913.
doi: 10.3389/fbuil.2025.1575913

Semantic and ontology-based analysis of regulatory documents for construction industry digitalization

Zarina Kabzhan[1], Alexandr Shakhnovich[1], Sergey Gorshkov[2], Yussuf Yemenov[1], Fedor Gorshkov[2] and Nazym Shogelova[1]*

[1]JSC Kazakh Research and Design Institute of Construction and Architecture, Almaty, Kazakhstan, [2]LLC Datavera, Almaty, Kazakhstan

In the context of the digitalization of the construction industry, the demand for automation in the analysis of regulatory documents is increasing. The significant volume, structural complexity, and frequent amendments of regulatory acts lead to semantic inconsistencies, duplication of provisions, and contradictions in requirements. The aim of this study is to develop a method for the automated analysis of regulatory documents in the construction sector based on the integration of ontological modeling and natural language processing (NLP) techniques. The relevance of the topic is driven by the need for the digital transformation of construction standardization, which involves eliminating redundant provisions, logical inconsistencies, and outdated references in the regulatory framework. The study proposes a methodology for constructing semantic "profiles" of regulatory statements, which include structured components: subject, predicate, object, modality, and additional conditions. A software prototype has been developed, implementing an algorithm for semantic matching of regulatory requirements using an ontological model that incorporates SKOS-based terminology, deontic logic, and domain-specific concepts of the construction sector. Experiments were conducted on a corpus of 14 regulatory documents of the Republic of Kazakhstan (≈242,000 words), demonstrating high computational efficiency (document analysis time <10 s) and acceptable quality metrics (F1-score up to 0.86). The results confirm the applicability of the proposed method for integration into automated regulatory compliance control systems and Building Information Modeling (BIM).

KEYWORDS

semantic analysis, ontology modeling, regulatory documents, digitalization of construction, NLP, automated analysis

1 Introduction

Modern construction is characterized by an extensive system of regulatory oversight throughout all stages of the asset life cycle—from design to operation. Projects and construction activities must comply with building codes, technical regulations, national, and industry standards that ensure safety, quality, and efficiency requirements. However, the large volume, complex structure, and frequent updates of the regulatory framework complicate the interpretation and practical application of requirements,

potentially hindering effective project design and implementation (Guo et al., 2021).One of the key challenges is the ambiguity of formulations resulting from discrepancies in terminology and definitions, which allows for multiple interpretations depending on the context of the document. Additionally, the duplication of provisions and logical contradictions among different regulatory acts further complicate the use of normative documents, making compliance verification more difficult and increasing the risk of legal conflicts (Guo et al., 2021). In this context, the automation of regulatory requirement analysis is viewed as a necessary component of the digital transformation of the construction industry, as it can enhance the transparency and speed of project approval processes.

Traditionally, compliance verification of construction projects with regulatory requirements is performed manually by experts, which is time-consuming, labor-intensive, and prone to subjective errors. For instance, the average review period for a commercial project in the city of Mesa (United States) is 18 working days, with analysis costs reaching up to $90 per hour (City of Mesa, 2012). Non-compliance, in turn, can lead to serious financial consequences—for example, Wal-Mart was once fined $1 million for violating environmental regulations (Salama and El-Gohary, 2013). Automating the compliance checking process could significantly reduce these costs and associated risks. Automated Compliance Checking (ACC) is currently viewed as a promising solution for accelerating expert review, reducing errors, and improving the efficiency of regulatory document management (Tang et al., 2012; Sydora and Stroulia, 2020). Over the past decades, various research approaches and software systems for ACC have been proposed, demonstrating the feasibility of partial automation. However, in practice, the level of automation in regulatory compliance within the construction domain remains low. To date, there has been no widespread adoption of digital methods in either the content of regulatory requirements or the procedures for their verification. This can be attributed to several limitations of the existing approaches (Zhang and El-Gohary, 2015).

Most modern ACC systems require manual rule authoring: experts manually extract requirements from normative texts and translate them into machine-readable formats (Zhang et al., 2022). This process is labor-intensive, prone to inaccuracies, and complicates system updates when regulations change. Many existing solutions rely on hardcoded rules or static databases, which limits their flexibility and scalability. Moreover, common methods for analyzing regulatory documents—such as keyword search and syntactic parsing—fail to capture semantic relationships and logical dependencies between provisions. This results in low accuracy of automated requirement interpretation, particularly for complex documents that contain numerous cross-references, exceptions, and conditional constructs. As experts have noted, full automation of compliance checking is not feasible without intelligent text processing—that is, transforming fragmented textual requirements into formal, machine-executable rules and linking them to digital project models (Zhong et al., 2012; Tierney, 2012). These limitations highlight a significant scientific and practical niche for developing novel methods to automate regulatory compliance control.

This study aims to develop a method for the automated semantic analysis of regulatory documents in the construction domain, based on ontological modeling and natural language processing (NLP) techniques, to overcome existing limitations in compliance automation. To achieve this goal, an analysis of current approaches to automated regulatory control is conducted to identify their weaknesses and justify the need for a new solution. The research involves the development of a multi-level ontological model designed to structure the content of regulatory documents and represent logical-semantic relationships between their provisions. In parallel, algorithms are developed to apply NLP methods for the automatic extraction of key terms and structured components of requirements from the text, as well as for generating their semantic representations ("profiles"). The proposed method enables the automatic identification of duplicated provisions, contradictions, and outdated references through the integration of the ontological model and semantic analysis. Its effectiveness is evaluated using a corpus of construction regulations, allowing for the assessment of inconsistency detection quality and demonstrating the method's potential for integration into digital platforms, including regulatory document management systems and Building Information Modeling (BIM) environments.

Achieving the stated goal and addressing the outlined objectives will enhance the degree of automation in the analysis of regulatory requirements, enable more accurate interpretation of normative documents, and provide a foundation for integrating compliance checking into BIM and other digital construction processes. The scientific contribution of this study lies in the advancement of methods for semantic analysis of regulatory texts, while its practical value is reflected in establishing the prerequisites for more efficient and reliable project expert review within the context of the industry's digital transformation.

2 Literature review

2.1 Analysis of regulatory requirements and ACC systems

The problem of Automated Compliance Checking (ACC) has been the focus of intensive research over the past several decades. As early as the 1980s and 1990s, initial attempts were made to formalize construction standards using expert systems and knowledge bases.

Existing ACC systems are based on transforming regulatory documents into logical rules suitable for machine analysis. Zhang and El-Gohary (2017) proposed a system that combines semantic text processing with logical reasoning, enabling the automatic extraction of requirements from building codes and their alignment with building BIM models (Zhang and El-Gohary 2017). Subsequently, the authors developed a method based on deep neural networks, which achieves high accuracy in extracting both syntactic and semantic elements from regulatory texts, reaching a precision of over 93% (Zhang and El-Gohary 2021).

The issue of scalability and flexibility of ACC systems when dealing with diverse regulatory documents remains a pressing challenge. In the study by Xue and Zhang, a mechanism was proposed for extending the set of rule transformation templates, which significantly improves the coverage of verifiable provisions and the accuracy of transformation through the iterative generation of logical expressions based on predefined patterns (Xue and Zhang, 2022).

Another emerging direction is the integration of large language models (LLMs), such as GPT-4, into the processes of automated regulatory compliance. Al-Turki et al. (2024) developed an LLM-based system that transforms building regulations into a structured YAML format using an active learning strategy. This approach significantly improves both the structural and semantic accuracy of interpreting regulatory requirements.

In addition to technical aspects, institutional prerequisites for the implementation of ACC systems are also being explored. Nama and Alalawi (2023) analyzed the adoption of automated compliance checking systems in Middle Eastern countries, identifying key stages in the transformation of building permit issuance processes and the integration of digital solutions into governmental workflows.

2.2 Digital technologies and BIM

Building Information Modeling (BIM) has become a key enabler in the development of Automated Compliance Checking (ACC) systems, providing a digital representation of the designed asset in the form of interconnected elements and their parameters. This data structure allows for the formalized description of design solutions, and consequently, enables programmatic verification of their compliance with regulatory requirements. In contrast to traditional checks based on drawings and textual documentation, the BIM-oriented approach improves accuracy, reduces time expenditures, and minimizes the impact of human error (Preidel and Borrmann, 2018).

Modern approaches to automated compliance checking involve the application of semantic processing to regulatory documents, followed by the alignment of extracted information with BIM model parameters. For example, in the study by Guo et al., a method was proposed in which regulatory requirements are extracted from texts using natural language processing (NLP), transformed into semantic rules, and then automatically applied to BIM models through the use of SPARQL queries (Guo et al., 2021).

More specialized applications of BIM for regulatory compliance checking have also demonstrated effectiveness. For instance, within the RegBIM project, an approach was developed to automate the verification of sustainable design solutions by extending the semantics of IFC to assess the environmental performance of projects (Kasim, 2015). In Pakistan, the implementation of a BIM-oriented automated checking system at the municipal level reduced the project documentation approval time from 1 week to 6 h, showcasing the high efficiency of digitalization in the expert review process (Aslam and Umar, 2023).

Literature reviews highlight that most modern systems rely on rule formalization based on the IFC standard. However, the interpretation of regulatory requirements remains the most labor-intensive stage: translating textual rules into machine-readable formats requires the use of either logical or ontological models, or specialized software plugins (Ismail et al., 2017).

Recent studies also propose innovative ACC system architectures. In particular, Li et al. (2024) developed a conceptual framework for automated compliance checking based on knowledge graphs, in which regulatory documents are transformed into ontologies using a Chinese NLP model and then applied to BIM models to detect violations. Another study introduced a system that leverages blockchain technology to ensure transparency and traceability of both automated and manual compliance verification stages within the BIM environment (Gao and Zhong, 2022).

2.3 Ontological modeling of regulatory knowledge

One of the most promising directions in the development of ACC systems is the use of ontological modeling for the formalization of regulatory knowledge. Ontologies enable the representation of building codes and standards as formal concepts, objects, properties, and logical relationships between them, thus making automatic interpretation and verification of regulatory requirements possible. This approach ensures a high degree of machine-readability and supports logical reasoning, which is essential for intelligent analysis of design documentation.

A number of studies focus on developing ontologies that capture the structure and semantics of building regulations. For example, Zhong et al. (2012) proposed an ontological approach to construction quality checking, based on the formalization of requirements using the OWL format and SWRL rules, which are applicable to inspection data. Jiang and colleagues developed a comprehensive multi-ontology merging method that enables the alignment of terminology between the building model and the regulatory ontology, as well as the application of logical reasoning through SPARQL queries (Jiang et al., 2022).

Of particular interest is the application of deontic logic—a logical system that describes concepts of obligation and prohibition—for the interpretation of regulatory requirements. Salama and El-Gohary (2011) proposed the use of deontological concepts in ACC systems to formalize rights and obligations within regulations, which is especially important for evaluating permissible and impermissible actions in construction.

One of the limitations of the ontological approach remains its labor intensity: the development of comprehensive ontologies requires the involvement of domain experts and significant time investment. Moreover, expressing dynamic or context-dependent requirements can be challenging within static ontological structures without the incorporation of additional logical mechanisms, such as first-order logic or conditional rules (Zhang and El-Gohary, 2017).

Nevertheless, the use of ontologies represents a powerful tool for integrating regulatory knowledge into intelligent systems, enabling a higher level of automation and reusability of formalized requirements across various tasks, including design, expert review, and compliance control.

2.4 Natural language processing (NLP) in regulatory analysis

Alongside ontological modeling, methods of Natural Language Processing (NLP) are actively evolving within the ACC domain, aiming to automate the analysis of regulatory document texts. The primary task of NLP in this context is to transform unstructured text—such as articles, clauses, and conditions of building codes—into formalized data suitable for subsequent logical processing and alignment with design model parameters.

In the early stages of such systems, template-based syntactic approaches were commonly used. Key markers (e.g., "shall," "is not permitted," "must") were employed to identify regulatory provisions and subsequently match them to elements of the construction model. However, these methods were limited in their ability to interpret complex constructions of regulatory language. In response to this, Zhang and El-Gohary (2016) proposed a semantics-oriented approach that applies a system of grammatical and semantic rules to extract objects, parameters, and logical relationships from regulatory texts.

Subsequently, hybrid algorithms were developed that combine syntactic parsing, semantic rules, and machine learning methods. In one study, the authors implemented a multi-component approach involving text classification, information extraction (IE), and the transformation of requirements into logical rules suitable for automated compliance checking. The system achieved high precision and recall in extracting quantitative requirements from international building codes (Zhang and El-Gohary 2013).

The specific characteristics of regulatory documents—such as complex syntactic structures, numerous conditions, exceptions, and cross-references—make the application of standard NLP models insufficiently effective. As a result, modern approaches combine linguistic rules with ontological information to eliminate semantic ambiguities and improve the accuracy of interpretation (Zheng et al., 2022).

A new and emerging direction involves the application of large language models (LLMs), such as GPT-4, to regulatory analysis tasks. Chen et al. proposed an architecture that combines LLMs with ontology to extract structured requirements from regulatory texts and verify them against BIM models. Pre-trained LLMs demonstrate a strong capability for interpreting complex texts and significantly reduce the need for manual data annotation (Chen et al., 2024).

However, the application of such models in the construction domain presents several challenges: the need to adapt to the specificity of professional language, limited transparency in logical reasoning, and high sensitivity to contextual nuances of regulatory phrasing. In this regard, the greatest potential lies in the integration of neural models with formalized knowledge in the form of ontologies and logical rules, which provides both flexibility and controllability in the interpretation of regulations.

2.5 Limitations of existing approaches and the research gap

Despite significant progress in the development of automated ACC systems, most existing solutions are focused on analyzing design models within the context of strictly predefined, manually formalized rules. These rules are typically encoded by domain experts based on regulatory texts and cover only limited aspects of requirements—such as geometric parameters, accessibility, fire safety clearances, and similar criteria (Jiang et al., 2022). As a result, ACC systems primarily operate as checklist-based verification modules and rarely treat the regulatory corpus itself as a subject of analysis.

Comprehensive semantic processing of the regulatory texts themselves—including the identification of internal contradictions, duplicated provisions, and outdated references—is currently addressed in only a limited number of studies. The vast majority of existing approaches assume that regulatory documents are consistent and free of contradictions, which does not always reflect real-world practice. In actual design conditions, regulations may contain ambiguous wording, overlapping or conflicting requirements, as well as references to outdated or repealed documents, all of which pose significant risks of misinterpretation (Zheng et al., 2022).

The analysis of regulatory overlaps and conflicts across multiple documents—such as different building codes, standards, and technical regulations—remains a largely unexplored area, despite its critical importance. Most modern ACC systems lack mechanisms for comparing provisions across various sources and do not track their currency, which necessitates manual verification and updating of references by domain experts (Beach et al., 2015).

Given these limitations, there is a growing need for approaches that can not only verify model compliance but also perform semantic analysis of the regulatory documents themselves. This includes automatic knowledge structuring, detection of logical inconsistencies, terminology alignment, and ensuring the relevance of regulatory provisions. One of the promising directions is the integration of ontological modeling with natural language processing (NLP) methods, enabling the extraction, formalization, and comparison of requirements from multiple sources (Zhang and El-Gohary 2016).

The approach proposed in this study is aimed precisely at addressing these challenges. Unlike conventional ACC systems focused solely on verifying BIM models, the developed method emphasizes in-depth digital processing of the regulatory texts themselves. By combining ontological knowledge representation with NLP techniques, the approach enables automation of regulation analysis at a more fundamental level, laying the groundwork for a fully digital chain: regulatory requirements–information model–compliance verification. Thus, the presented research fills the identified scientific and practical gap, advancing the construction industry toward a new stage of digitalization that enhances both the efficiency and quality of design decisions.

3 Methods and materials

For the purpose of formalization and automated analysis of regulatory requirements applied in the construction industry, this study proposes a methodology based on the use of ontological modeling, natural language processing (NLP) techniques, and semantic analysis (Chen et al., 2024). The domain of interest is the set of building codes and regulations adopted in the Republic of Kazakhstan. The study is grounded in the following key principles.

First, regulatory documents possess a complex, multi-level structure that includes nested conditions, exceptions, cross-references, and hierarchically organized requirements.

Accordingly, the proposed methodology incorporates a step-by-step analysis of the syntactic and semantic characteristics of regulatory statements, aimed at their subsequent formalization and alignment.

At the first stage, preliminary linguistic processing of the texts is performed, including tokenization, lemmatization, part-of-speech (POS) tagging, dependency parsing, and coreference resolution.

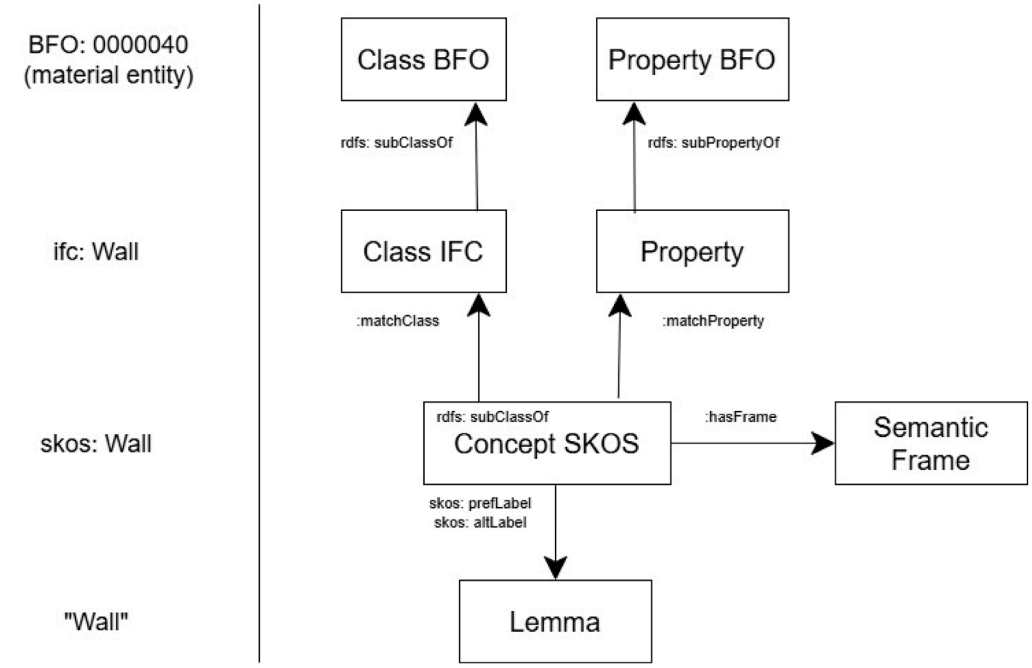

FIGURE 1
Relationship between elements of the proposed ontologies.

These procedures are carried out using the DataVera EKG Language Processing (EKG LP) software module (DataVera, 2025), which is built on the SpaCy library and adapted to the specifics of regulatory vocabulary.

At the second stage, textual fragments are aligned with the ontological model, which is represented as a set of interconnected ontologies (Figure 1):

- Upper-level ontology (based on BFO), used to represent universal categories such as objects, processes, and relationships;
- Domain ontology of the construction sector (based on IFC), covering capital construction assets, engineering systems, and life cycle processes;
- Regulatory statement ontology, based on deontic logic, describing the structure of norms (subject, modality, action, object, and applicability condition);
- Terminology ontology (SKOS model), providing linkage between the terms used in regulatory documents and the concepts of the domain ontology.

The formalized representation of regulatory provisions is carried out in the form of semantic profiles, which include the following elements: subject (addressee of the requirement), modality (obligation, possibility, prohibition), predicate (action or characteristic), object (result of the action), as well as additional attributes (conditions, exceptions, time frames, etc.).

To account for the complex structure of regulatory texts, the methodology implements mechanisms for:

- Detection of nested conditions (through the analysis of syntactic structures and conditional operators);
- Processing of exceptions, formed through negation constructs or limitations on the scope of regulations;
- Reconstruction of hierarchical relationships between regulatory provisions, using structural markers and contextual analysis of headings, articles, and subsections.

At the final stage, a comparative semantic analysis is performed, aimed at identifying:

- Duplicated provisions (when key elements of the semantic profile match);
- Contradictions (when there are discrepancies in modalities or conditions of application);
- Semantic inconsistencies (in definitions of terms and interpretations of concepts).

The comparison of semantic profiles is carried out based on a calculated similarity metric, the threshold value of which is determined empirically. In the case of significant discrepancies, the corresponding fragments are forwarded for expert review.

The developed system is designed for the automated semantic analysis of regulatory documents, identifying contradictions, duplicated provisions, and semantic inconsistencies. The architectural solution (Figure 2) is based on the use of ontological models, graph and relational databases, as well as natural language processing (NLP) methods.

The system includes several key components that ensure its functionality. A graph-based RDF triple store database (Apache Fuseki) is used for storing ontological models, enabling complex semantic queries and analysis of relationships between concepts. A relational or document-oriented storage system (PostgreSQL) is employed to store the results of the linguistic analysis of regulatory

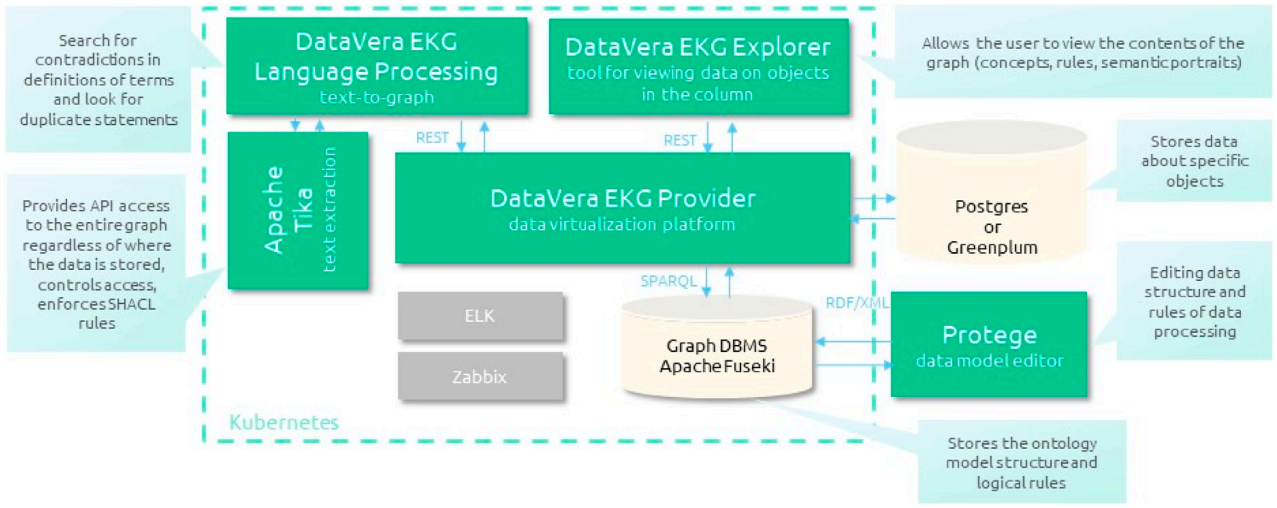

FIGURE 2
Architecture of the automated system for processing regulatory document texts.

texts (Jadala and Burugari, 2024). An important element is the data management platform (DataVera EKG Provider (DataVera, 2025)), which ensures information storage in accordance with the ontological model, supports both synchronous and asynchronous APIs, executes SPARQL queries, and performs data validation using SHACL rules (Ke et al., 2024). The system also includes application software modules, such as the linguistic analysis module for regulatory documents (DataVera EKG LP (DataVera, 2025)) and the semantic analysis module, which identifies contradictions in terminology and detects duplicated provisions. Monitoring and logging tools, such as ELK and Zabbix, are used to ensure system oversight and log collection (Bilobrovets, 2023).

The system is implemented as a set of containers deployed in a Kubernetes environment (Poniszewska-Marańda and Czechowska, 2021), which ensures its scalability and fault tolerance.

The processing of regulatory texts is performed in stages, starting with grammatical and semantic analysis (DataVera, 2025):

- Sentence structure analysis includes POS-tagging and dependency parsing, which allows for the identification of parts of speech and the establishment of grammatical dependencies between words. Coreference resolution is also performed, involving the substitution of nouns for pronouns and clarification of implied elements in the statement.
- Lemmatization ensures the conversion of word forms to their base form, simplifying subsequent processing and matching.
- Semantic matching involves identifying the concepts corresponding to the words in the sentence based on ontological models. In the absence of an exact match in the existing ontology, the system automatically generates *ad hoc* concepts limited to the specific context of the document.
- Formation of the semantic profile involves identifying subjects, predicates, modalities, objects, circumstances, and other elements necessary for the structured representation of regulatory content.

The result of the algorithm's operation is the formalized representation of each statement in the form of a set of semantic profiles, suitable for further analysis. Based on the obtained semantic profiles, a comparison of regulatory provisions is performed, allowing for the identification of contradictions, duplication, and semantic inconsistencies.

The identification of contradictions in terminology is carried out by analyzing statements that contain definitions of regulatory terms. The comparison of such statements allows for classifying the results into three groups (Liu et al., 2020):

- Semantic equivalence (the definitions are identical or close in meaning).
- Difference in scope (one definition is a specific case of the other).
- Semantic contradiction, when mutually exclusive interpretations of the same term are identified.

The search for duplicated regulatory provisions is performed by comparing the key elements of the semantic profile. If statements from different documents have matching predicates, objects, subjects, modalities, and additional parameters, the system calculates a numerical similarity metric. If the threshold value is exceeded, the statements are considered duplicated.

Similarly, contradictory statements are identified. If two statements refer to the same entity (matching subject, predicate, and object) but have different modalities, a logical contradiction is detected. In cases where additional elements of the semantic description differ, the inconsistency is evaluated quantitatively. If the discrepancy exceeds the established threshold, the divergences are forwarded for expert analysis.

The developed method for analyzing regulatory documents has a number of limitations related to the depth of semantic processing. First, the system evaluates the semantic profile of each statement in isolation, which excludes the possibility of analyzing situations where a single statement in one document corresponds to multiple statements in another. Second, the current implementation does not

account for the temporal aspect of regulatory provisions, meaning it does not analyze to which time period a particular directive applies (past, present, or likely future). Third, the system does not generate a comprehensive semantic description of the situations to which the requirements apply, but is limited to representing the regulatory directive in a structured form. While this simplifies the development and implementation of the system, such a level of formalization is insufficient for automated compliance checking and is intended solely for identifying inconsistencies and duplications in regulatory provisions.

To address the identified limitations, it is proposed to further develop the methodology across several interrelated directions. One of the key vectors is the development of a mechanism for inter-document semantic aggregation, which would enable the establishment of relationships such as equivalence, specification, logical entailment, and subordination between regulatory statements—both within a single document and across multiple sources. This would allow for the modeling of complex regulatory dependencies and improve the accuracy of contradiction detection.

Special attention is planned to be given to incorporating the temporal aspect of regulatory requirements. This involves annotating regulatory provisions with temporal markers (such as effective date, duration, and period of applicability), followed by integration with temporal ontologies. Such an approach will enable the tracking of regulatory evolution and the assessment of the applicability of provisions at a given point in time.

Another important direction is the modeling of regulatory situations through the expansion of the ontological model by incorporating concepts that describe typical scenarios for the application of requirements. This creates a foundation for shifting from the analysis of isolated provisions to a comprehensive assessment of regulatory conditions based on the context of design or operation of built assets. Such a level of detail will enhance the practical relevance of the developed system in professional practice.

To improve the completeness and validity of the analysis, it is proposed to integrate logic-based semantic reasoning using ontological rule languages such as SHACL or SWRL. This will enable not only the interpretation of individual statements, but also the formalization of logical relationships between them, thereby allowing for deductive consistency checking of regulatory requirements.

Finally, an important element of future work is the implementation of a contextual semantic disambiguation mechanism using trainable language models (e.g., BERT or GPT) adapted to a corpus of regulatory texts. The use of such models will enable accurate interpretation of terms and constructions depending on their usage context, especially in cases where the same concept may have different meanings in different sections or documents.

The implementation of the proposed directions will eliminate current limitations and significantly expand the functional capabilities of the system. This will pave the way for the development of a full-featured intelligent platform for regulatory analysis, capable of supporting tasks related to design, expert review, auditing, and legal compliance in the context of the construction industry's digital transformation.

The proposed architecture and methodology enable effective analysis of regulatory documents in the construction sector by providing their structured representation, identifying semantic inconsistencies, and supporting the development of a more coherent regulatory framework.

4 Results

To assess the applicability of the proposed approach, the study employed the EKG LP software suite, developed to address a wide range of text processing tasks. The choice of this software is justified by its ability not only to extract key entities and relationships from text, but also to generate an ontological representation of document structure, which is critically important for analyzing complex regulatory acts. Unlike many other systems, EKG LP provides built-in tools for constructing knowledge graphs and performing semantic annotation, enabling the automation of regulatory requirement interpretation, contradiction detection, and the formalization of logical relationships between provisions.

In addition, the software suite is integrated with corporate databases and electronic document management systems, making it particularly valuable in the context of digital transformation in the construction industry. Although EKG LP has not yet achieved widespread adoption among construction professionals, its potential is actively being explored within projects aimed at the digitalization of the regulatory and technical framework, including initiatives for implementing information modeling technologies and developing digital codes and standards. The present study demonstrates the applicability of this tool specifically in the context of construction regulation tasks, confirming its relevance and effectiveness within this domain.

The first stage of text processing in EKG LP involves lemmatization and grammatical structure parsing of sentences, performed using tools from the SpaCy framework (Díaz et al., 2024). During analysis, each element of the text is assigned morphological and syntactic characteristics, and the identified grammatical dependencies are structured hierarchically. These dependencies are visualized using the displacy tool and are shown in Figure 3. The output of this processing phase, printed from the EKG LP source code, is presented in Figure 4. As an example, consider the sentence: "Joint connections of prefabricated elements and multilayer structures shall be designed to withstand temperature deformations and forces arising from uneven foundation settlement and other operational impacts."

The result of this processing phase is a structured representation of the sentence, in which each word and punctuation mark is linked to its lemmatized form along with its grammatical function. On the left side of the visualized representation, the words of the original sentence are arranged according to the identified syntactic dependencies. On the right side of the table (Figure 4), each word is annotated with its part of speech (POS-tag) and the type of syntactic relation it holds with other sentence elements (Relation type), enabling further processing at the level of semantic dependencies.

Based on the data obtained, EKG LP constructs a "semantic profile" of the statement (Table 1), the structure of which is analogous to the model used in the Nòmos two framework (Mandal et al., 2015). During this process, the core semantic structure of the text is identified, including the

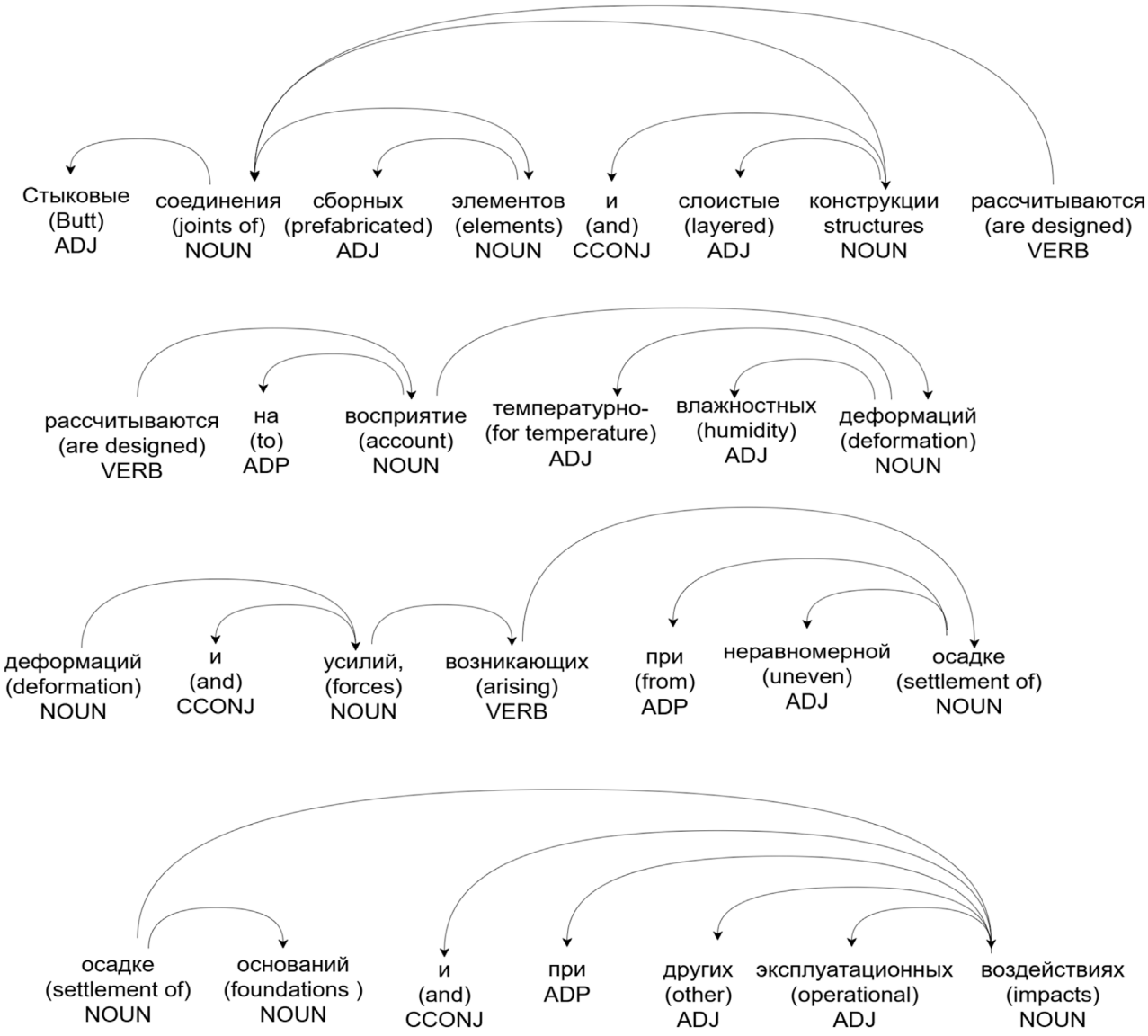

FIGURE 3
Result of lemmatization and grammatical structure parsing of the sentence.

key components of the statement: predicate, object, subject, and modality.

To illustrate, consider the analysis of a specific example, where in the phrase "connections are designed to withstand" the following semantic components are extracted: "connection" as the subject, "designed" as the predicate (normalized to the base form "design"), and "withstand" as the object.

In addition, dependency chains are generated for both the subject and the object, enabling a more detailed description of regulatory provisions and contributing to the precise identification of their semantic structure. The resulting semantic structures are subsequently used to detect contradictions, duplicated provisions, and semantic inconsistencies in regulatory documents.

To verify the proposed approach, the EKG LP software suite was used, developed for the automated analysis of regulatory documents. Its primary purpose in this study is to identify duplicated requirements, analyze the semantic similarity of phrases, and detect contradictions in regulatory provisions.

In its default configuration, EKG LP generates a "semantic profile" for each statement, consisting of seven key components: subject, predicate, object, modality, negation, definition, and complement/circumstance. The analysis revealed that this structure is sufficient for accurately representing simple sentences; however, regulatory documents in the construction sector are characterized by a high degree of syntactic complexity. As a result, the basic algorithm requires further refinement to enable more accurate modeling of complex statements. Nevertheless, even the current version of the algorithm demonstrates satisfactory performance in comparing phrases with similar semantic structures.

When analyzing two semantically similar statements (Martinez-Gil and Chaves-González, 2022), EKG LP generates their semantic profiles, which turn out to be nearly identical, with only minor

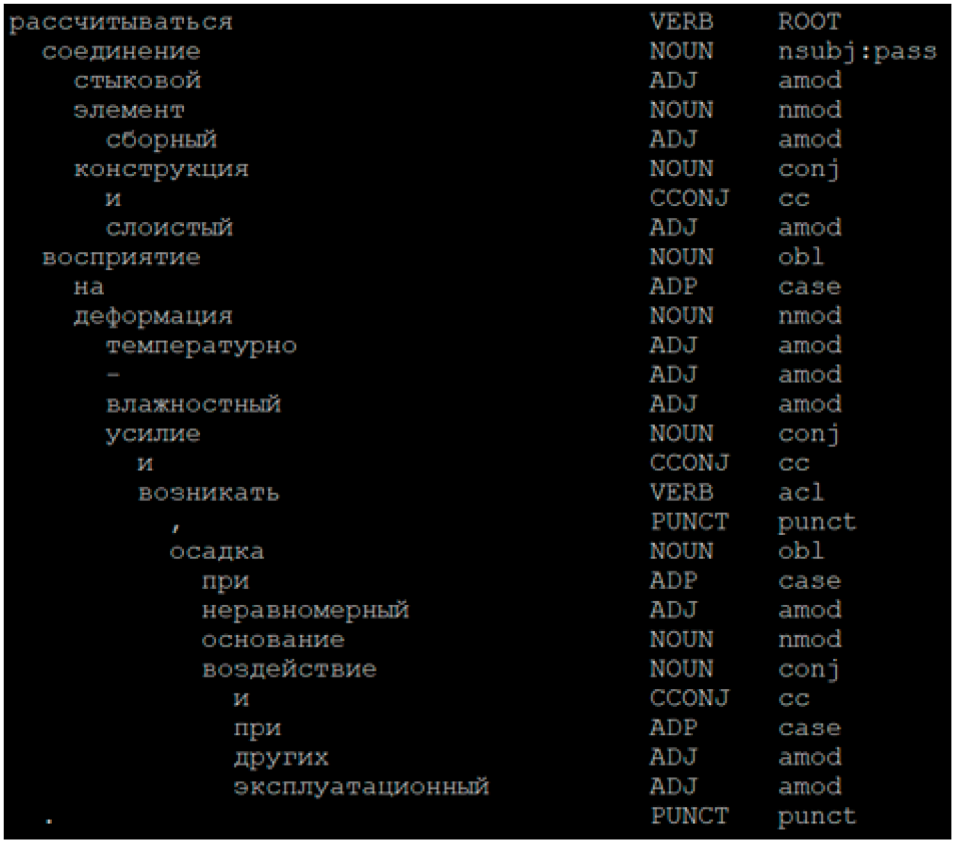

FIGURE 4
Output of lemmatization and grammatical structure parsing of the sentence.

TABLE 1 Comparison of «semantic portraits».

Subject	Connection	Connection
Predicate	Calculate	Calculate
Object	Perception	Perception
Modality	-	-
Negation	-	-
Definition	Joint, prefabricated, layered, temperature, moisture, irregular, others, operational	Joint, prefabricated, multilayer, thermal, irregular, others, operational

differences in definitions. The software calculates a semantic similarity metric ranging from −1 to 1, where −1 indicates completely opposite meanings and one indicates full equivalence. In the example considered, the metric value was 0.91, indicating a high degree of similarity between the phrases. By setting a threshold for this metric, it becomes possible to identify regulatory requirements that are duplicated either within a single document or across different regulatory sources. This confirms the applicability of the proposed methodology for the automated detection of redundant regulatory information (Colla et al., 2020).

When comparing semantic profiles, statements are considered equivalent only if they share the same predicate. Otherwise, the comparison result is set to zero. Negative metric values may occur in cases where the analyzed phrases differ in modality (e.g., *"may"* vs. *"shall"*) or when one of the statements includes predicate negation (e.g., *"is designed"* vs. *"is not designed"*).

One of the limitations of the basic algorithm is that it does not account for the semantic similarity of individual lexemes. As a result, phrases that are equivalent in meaning but differ in lexical composition may receive a semantic similarity score of zero. To address this issue, two possible approaches can be considered:

- Using vector-based models (e.g., Word2Vec, BERT), which enable the assessment of term similarity based on their contextual usage. However, for domain-specific texts, such models often demonstrate limited accuracy, as construction-related terms tend to be semantically close to each other, reducing the algorithm's discriminative capability.
- Shifting from lemmata to concepts using a SKOS-based ontological model, which makes it possible to account for hierarchical relationships between terms, such as equivalence, broader terms, and narrower terms. This approach enables more accurate computation of semantic similarity and allows for the identification of logical contradictions at a deeper level.

TABLE 2 Semantic portraits of phrases constructed from lemmas.

Elements of the semantic portrait	Proposition 1	Proposition 2
Subject	-	-
Predicate	Consider	Observe
Object	Process, purpose	Governance, interest
Modality	Need	Necessary
Negation	-	-
Definition	-	-
Complement/ Circumstance	Management, information, object, owner	Data, asset, owner

To demonstrate the advantages of using a conceptual model, let us consider two synthetic phrases:

- Sentence 1: "In the process of managing information about assets, it is necessary to consider the goals of their owners."
- Sentence 2: "During asset data management, the interests of their holders must be respected."

Despite the semantic equivalence of these statements, their lexical composition differs, which leads the lemma-based algorithm to assign them a semantic similarity score of zero (Table 2).

To overcome this limitation, a SKOS-based conceptual model was developed, incorporating the following terminological relationships:

- *"Consider"* = *"Respect"* (equivalent terms)
- *"Must"* = *"Necessary"* (equivalent terms)
- *"Holder"* = *"Owner"* (equivalent terms)
- *"Goal"* < *"Interest"* (narrower term)
- *"Information"* > *"Data"* (broader term)
- *"Asset"* < *"Object"* (narrower term)

This ontology was deployed in the Apache Fuseki system (Figure 5), which is accessed by EKG LP. During the processing of semantic profiles, the algorithm replaces lemmata with their corresponding concepts (Table 3), allowing for a more accurate calculation of similarity.

As a result of recalculation, the semantic similarity metric for the considered phrases was 0.48, reflecting their partial equivalence. In this approach, terms with broader or narrower meanings are interpreted as 75% matches, which enables the adjustment of the metric calculation algorithm accordingly.

To assess the scalability of the proposed approach, a preliminary evaluation of the lexical core of regulatory documents in the construction sector was conducted. The analysis was performed using Apache Tika to extract text from 14 regulatory documents provided by the client. Only Russian-language text was processed.

The sample used for the experimental evaluation comprised 14 regulatory documents with a total length of approximately 242,000 words, which is equivalent to an average industry-level regulatory corpus. Despite the representativeness of the content (the documents cover various aspects of construction regulation—design, operation, information modeling, etc.), this volume should be considered a pilot dataset suitable for initial testing of the proposed method's effectiveness.

From the standpoint of scalability, the evaluation conducted on this dataset made it possible to identify key characteristics of the algorithm's performance and confirm its applicability to real regulatory data. However, to ensure a high degree of generalizability and robustness of the method against variability in phrasing, structure, and lexical patterns, further expansion of the corpus is required.

Expanding the size of the training and test document sets, including a broader range of regulations (such as international standards, technical regulations, sanitary and fire safety codes), as well as covering documents with varying structural complexity, will enhance the validity of the obtained quality metrics. An extended corpus will make it possible to more accurately calibrate the parameters of semantic similarity, test the algorithms across a wider variety of contexts, and identify potential bottlenecks in the ontological model.

Thus, the effect of scaling lies not only in improving the reliability of the evaluation, but also in enhancing the ability of the developed method to adapt to new types of regulatory texts—an aspect that is critically important for its future practical application in the context of a constantly evolving regulatory landscape.

Based on lemmatization performed using Pymorphy2, the following quantitative characteristics were obtained:

- Number of unique lemmata: 9,400
- Percentage coverage by frequent lemmata: 91% of the total document text

A significant portion of regulatory documents employs a relatively limited set of key terms, which makes the task of constructing a SKOS-based conceptual model feasible.

Additionally, an analysis of the most frequently occurring words in the examined documents was conducted (Table 4). The ten most common lemmata account for 9% of the total text volume, indicating a high degree of lexical unification in regulatory documents. This observation supports the feasibility of effective conceptualization of industry-specific terminology, including the establishment of semantic frames and dependencies.

The analysis demonstrated that the use of semantic profiles in combination with a SKOS-based ontological model enables effective identification of duplicated regulatory requirements and assessment of their semantic similarity. The developed method also supports the detection of logical contradictions at the level of terms and their relationships.

The introduction of an ontological model in place of simple lemma comparison represents a fundamentally different level of text analysis. While lemmatization provides only superficial matching of word forms, the ontological model allows for the consideration of hierarchical and associative relationships between terms, their contextual roles, and their affiliation with specific concepts. This approach enables a deeper and more contextually grounded understanding of regulatory texts, which is critically important for the automated interpretation of requirements and the identification of logical relationships between document provisions.

```
12 SELECT * WHERE { skos:Information ?prop ?value }
13
```

Table | Response | 5 results in 0.163 seconds | Simple view ☐ Ellipse ☑ | Filter query results | Page size: 50

	prop	value
1	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/2004/02/skos/core#Concept>
2	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/2002/07/owl#NamedIndividual>
3	<http://www.w3.org/2004/02/skos/core#prefLabel>	"информация"@ru
4	<http://www.w3.org/2004/02/skos/core#narrower>	<http://www.w3.org/2004/02/skos/core#Data>
5	<http://www.w3.org/2000/01/rdf-schema#label>	"информация"@ru

FIGURE 5
Concept "Information" in the results of a SPARQL query to Apache Fuseki

TABLE 3 Substitution of lemmas for concepts in the "semantic portrait" of the EKG LP.

Elements of the semantic portrait	Proposition 1	Proposition 2
Subject	-	-
Predicate	skos:Consider	skos:Consider
Object	Process, skos:Goal	Management, skos:Interest
Modality	skos:Need	skos:Need
Negation	-	-
Definition	-	-
Complement/ Circumstance	Management, skos:Information, skos:Object, skos:Owner	skos:Data, skos:Asset, skos:Owner

TABLE 4 Ten most frequently occurring words and their frequency of use (excluding prepositions).

Terms	Quantity
Object	3 151
Construction	2,469
Work	2,455
Construction	2,441
Project	2,329
Informational	2,218
Building	2,100
Information	1,964
Model	1,787
System	1,685

The results obtained in the course of the study confirm that the construction of a conceptual (ontological) model of industry-specific terminology is a labor-intensive but feasible process that can significantly enhance the accuracy and completeness of automated analysis of regulatory documents in the construction sector.

5 Discussion

5.1 Analysis of identical and similar terminological definitions

To validate the proposed algorithm using practical examples, a series of experiments was conducted to identify duplicated and similar regulatory provisions in construction regulations.

The following documents were used as test materials:

– SP RK 1.02-120-2019 *"Application of Information Modeling in Construction Organizations"* (Zakon.kz, 2025);

– SP RK 1.02-121-2019 *"Application of Information Modeling in Operating Organizations"* (Zakon.kz, 2025);

These documents contain a significant number of identical or similar definitions and provisions, making them well-suited for analyzing the capabilities of the developed method.

Before conducting semantic analysis, the texts underwent preliminary processing aimed at removing elements that hinder the automated parsing of regulatory documents. At this stage, textual data were extracted from the original PDF files using the Apache Tika tool (Burgess and Mattmann, 2014), allowing them to be obtained in a structured format. Subsequently, auxiliary elements of the documents—including headers, page numbers, line breaks, titles, and other components not affecting the semantic content—were removed. After cleaning, the data were processed in the EKG LP environment, resulting in sets of semantic profiles of statements. The final processing step involved comparing the obtained semantic structures of the two documents at the lemma level, without applying the conceptual model.

The EKG LP algorithm successfully identifies matching definitions present in both documents. For example, during the analysis, it correctly detected a duplicated definition:

> "Stakeholder: A person, group, or organization that can affect, be affected by, or perceives itself to be affected by decisions, activities, or outcomes of a project."

However, due to the specifics of the SpaCy framework, variations in the grammatical structure analysis of the same phrase may occur. As a result, the similarity metric for comparable definitions does not always reach 100%. To improve accuracy, the implementation of an additional post-processing algorithm is proposed, which would take into account the sequence and set of words in the sentence. This would allow for more precise determination of textual identity.

5.2 Identification of similar statements in the text

In addition to terminological definitions, the algorithm also detects similar statements appearing in both documents. For example,: *"The information management function includes monitoring compliance with standards and requirements (the organizational standard for building information modeling technology, asset information requirements), monitoring the content and updating of the asset information model (AIM), and ensuring adherence to information approval and coordination procedures."*

The developed method demonstrated high computational efficiency. The following performance indicators were obtained during the experiments:

- Document preparation and grammatical structure parsing take less than 10 s per document.
- Comparison of statements between two documents is performed in less than 1 s.

Thanks to the high processing speed, it becomes feasible to implement a method for large-scale document comparison. In particular, a database of grammatical parsing results for regulatory acts can be created, enabling pairwise comparison of each document with all others to automatically identify duplicated and contradictory provisions.

The developed method for analyzing regulatory documents enables the task of automatic contradiction detection based on the comparison of semantic profiles of statements. In particular, inconsistencies may arise from differences in numerical values, mismatches in modality, or the presence of negation.

To illustrate, consider the following regulatory provision:

> "Tactile indicators serving a warning function on pedestrian pathway surfaces shall be placed no less than 0.8 m / 0.6 m from the information object or the beginning of a hazardous area, change in direction, entrance, etc."

In this case, the semantic profiles of phrases containing numerical characteristics will be nearly identical, with the only source of discrepancy being the difference in numerical values. Since the sentence structure parser marks such values as POS = NUM and dep = nummod, their comparison does not present technical difficulties and can be implemented as a specialized application module.

Another common type of contradiction is a difference in modalities. Consider the following examples:

- *"When designing the site, it is necessary to take into account the condition of natural landscape development."*
- *"When designing the site, it is recommended to take into account the condition of natural landscape development."*

Despite the similarity in the overall structure of the sentences, their semantic profiles differ due to the use of different modal operators: *"necessary"* and *"recommended."* During analysis, this results in a semantic similarity metric value of −0.98, indicating a high degree of semantic divergence between the statements. A similar result would be obtained in the case of opposing modalities (e.g., *"shall"*/ *"shall not"*).

The contradiction detection method is applicable to at least three types of discrepancies:

- Differences in numerical values within regulatory provisions;
- Mismatches in the modality of statements;
- Presence of negation that alters the meaning of the statement.

It should be noted that, despite the high level of automation, human intervention remains necessary at certain stages to improve the accuracy and reliability of the results. In particular, expert review helps to:

- Interpret the context of statements not captured by the algorithm;
- Clarify cases of terminological discrepancies related to industry-specific language;
- Determine the criticality of the identified inconsistencies.

Combining automated analysis with expert evaluation ensures more reliable and well-grounded detection of contradictions in regulatory documents. To effectively detect such inconsistencies, the algorithm uses a semantic similarity threshold calibrated on real-world data. For example, if the similarity of semantic profiles exceeds 60% but one of the listed discrepancies is detected, the statements are considered potentially contradictory and are flagged for further review.

5.3 Comparative analysis with existing methods

To assess the contribution of the developed method, a comparative analysis was conducted, evaluating its characteristics against other contemporary approaches used for automated analysis of regulatory documents. The following baseline solutions were selected:

- Text matching method based on TF-IDF and Jaccard similarity metric (Plansangket and Gan, 2015);
- Topic modeling using Latent Dirichlet allocation (LDA) (Subramanian et al., 2024);
- Sentence-BERT model, representing a modern class of transformer-based models for semantic similarity estimation (Westin, 2024);

TABLE 5 Comparative summary of methods.

No	Method	Precision	Recall	F1-score	Interpretability
1	TF-IDF + Jaccard	0,68	0,53	0,6	0,3
2	LDA (Topic modeling)	0,61	0,49	0,54	0,4
3	Sentence-BERT	0,87	0,8	0,84	0,6
4	Proposed method	0,89	0,84	0,86	0,9

– The proposed ontological method, which uses the formalization of statements in the form of semantic profiles and ontological relationships (Motta and Ladouceur, 2017).

The comparison was carried out using the following criteria (Table 5):

– Precision — the proportion of correctly classified duplicates and contradictions among all identified by the system;
– Recall — the proportion of actual duplicates and contradictions correctly detected by the system;
– F1-score — the harmonic mean of precision and recall;
– Interpretability — expert evaluation of the transparency of the algorithm's logic and the interpretability of its results (on a scale from 0 to 1).

The comparative analysis conducted showed that the proposed ontological method for semantic analysis of regulatory documents demonstrates high efficiency in identifying duplicated and contradictory provisions. According to standard quality metrics (precision = 0.89, recall = 0.84, F1-score = 0.86), the method is comparable to or outperforms existing solutions, including models based on Sentence-BERT, while having significantly higher interpretability (0.9 on the expert scale).

Unlike general-purpose text processing methods, the proposed methodology takes into account the specifics of regulatory documents: the presence of modal constructions, logical constraints, the subject structure of requirements, and the hierarchy of statements. The use of a semantically rich format for representing regulatory provisions in the form of "profiles" with ontological annotations not only enables the identification of semantic discrepancies but also provides a foundation for the automated logical verification of the consistency of regulatory requirements.

Thus, the developed approach can serve as the foundation for building intelligent systems for regulatory analysis support, providing both high-quality metrics and transparency in decision-making.

The comparison results confirm the relevance and practical significance of the proposed methodology in the context of the digitalization of the construction industry and the reform of the regulatory framework.

5.4 Practical recommendations for applying the developed method

To integrate the developed approach into regulatory analysis processes, project documentation expertise, and regulatory framework management in the construction industry, the following aspects of practical application should be considered:

– Use at the regulatory expertise stage. The method can be implemented as a supplementary tool in the regulatory document review process—to automatically identify duplicated and contradictory provisions between existing and proposed regulations. This is particularly relevant when developing new versions of standards, technical regulations, and departmental regulations.
– Support for the digital transformation of the regulatory framework. The proposed approach can be utilized within the digitalization of the regulatory and technical framework, including the construction of ontologically organized databases of building codes and their automated verification. This creates the foundation for transitioning from textual representation of requirements to their formalized, machine-readable structure.
– Integration into project documentation information systems (BIM). Integrating the developed method into software systems supporting building information modeling (BIM) technologies will allow for automatic verification of design decisions against current regulations at the early stages of design, helping to prevent regulatory conflicts. This is especially useful when generating automatic compliance reports for model requirements.
– Training and preparation of experts. To ensure effective implementation, it is recommended to develop training modules for specialists in technical standardization, expertise, and design, explaining the logic behind semantic profile construction, the principles of ontology formation, and the interpretation of analysis results.
– Support for the development of new regulations. The method can be applied during the regulatory drafting stage to compare draft documents with existing regulations, assess the consistency of provisions, and ensure the uniformity of terminology, particularly in cases where documents of different levels (state, industry, corporate) are functioning simultaneously.

The proposed methodology has a high degree of adaptability and can be integrated into the practices of various participants in the construction process—from regulatory bodies and expert centers to design and operational organizations. Implementing this approach will enhance the consistency of the regulatory framework, reduce the risks of regulatory contradictions, and support a more sustainable digital transformation of the construction industry.

6 Conclusion

This work proposes and validates a method for the automated analysis of regulatory documents in the construction industry, based on a combination of natural language processing (NLP) techniques and ontological modeling. The developed algorithm ensures the formation of semantic profiles for regulatory provisions, identification of duplicated and contradictory statements, and verification of the relevance of referenced documents.

The use of an ontocentric approach allows for the formalization of knowledge contained in regulatory documents and its integration into digital platforms for managing regulatory requirements. The developed methodology demonstrated its effectiveness through the analysis of the building codes of the Republic of Kazakhstan, showing its ability to identify logical inconsistencies, automate the classification of regulatory provisions, and compare requirements across different documents.

Experimental studies have confirmed the high computational efficiency of the developed algorithm, making it suitable for use in scalable regulatory document analysis systems. In particular, the processing speed of a single document does not exceed 10 s, and the comparison of regulatory provisions is completed in less than 1 s. This enables the implementation of a concept for large-scale comparative analysis of regulatory documents, identifying inconsistencies across large datasets of legal information.

Despite the achieved results, the algorithm requires further improvement. Key directions for future research include:

- Expanding the system's functionality to recognize the type of regulatory statements (definitions, requirements, notes, etc.);
- Automatic extraction of document structure (sections, articles, tables) to improve the processing of complex regulatory acts;
- Implementing semantic disambiguation algorithms using large language models (LLMs) to enhance text analysis accuracy;
- Integrating the system into the BIM ecosystem to ensure automated compliance control of design decisions with regulatory requirements.

The results of the study confirm that the use of ontological modeling combined with NLP methods is a promising direction for the automated analysis of regulatory documents. The developed method could serve as the foundation for creating intelligent Automated Compliance Checking (ACC) systems, supporting the digitalization of the construction industry and enhancing the efficiency of regulatory governance.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ZK: Conceptualization, Validation, Writing – review and editing. AS: Conceptualization, Project administration, Writing – review and editing. SG: Investigation, Methodology, Writing – original draft. YY: Data curation, Visualization, Writing – original draft. FG: Software, Writing – original draft. NS: Formal Analysis, Resources, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The methods for creating "semantic profiles" were studied as part of the R&D project "Research on Methods for Automating Semantic Control of Term Definitions in Regulatory Documents of the Construction Industry" (JSC "KazNIISA," LLP "DataVera").

Conflict of interest

Authors SG and FG were employed by LLC Datavera.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Correction Note

A correction has been made to this article. Details can be found at: 10.3389/fbuil.2025.1624950.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Turki, D., Hettiarachchi, H., Gaber, M., Abdelsamea, M., Basurra, S., Iranmanesh, S., et al. (2024). Human-in-the-Loop learning with LLMs for efficient RASE tagging in building compliance regulations. *IEEE Access* 12, 185291–185306. doi:10.1109/ACCESS.2024.3512434

Aslam, B., and Umar, T. (2023). Automated code compliance checking through building information modelling. *Proc. Institution Civ. Eng. - Struct. Build.* 177, 822–837. doi:10.1680/jstbu.22.00214

Beach, T., Rezgui, Y., Li, H., and Kasim, T. (2015). A rule-based semantic approach for automated regulatory compliance in the construction sector. *Expert Syst. Appl.* 42, 5219–5231. doi:10.1016/j.eswa.2015.02.029

Bilobrovets, I. (2023). Network threat detection technology using Zabbix software. *Mod. Inf. Secur.* 54. doi:10.31673/2409-7292.2023.020003

Burgess, A., and Mattmann, C. (2014). "Automatically classifying and interpreting polar datasets with Apache Tika," in Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), 13-15 August 2014, Redwood City, CA, USA. 863–867. doi:10.1109/IRI.2014.7051982

Chen, N., Lin, X., Jiang, H., and An, Y. (2024). Automated building information modeling compliance check through a large language model combined with deep learning and ontology. *Buildings* 14 (7), 1983. doi:10.3390/buildings14071983

City of Mesa (2012). Construction plan review. *Official Website City Mesa, Ariz.* Available online at: http://www.mesaaz.gov/devsustain/PlanReview.aspx (November 25, 2022).

Colla, D., Mensa, E., and Radicioni, D. (2020). Novel metrics for computing semantic similarity with sense embeddings. *Knowl. Based Syst.* 206, 106346. doi:10.1016/j.knosys.2020.106346

DataVera (2025). DataVera NLU: natural language understanding platform. Available online at: https://datavera.kz/ru/nlu.html February 11, 2025).

Díaz, H., Braumann, S., Van Der Poel, J., Gog, T., and De Bruin, A. (2024). Towards adaptive support for self-regulated learning of causal relations: evaluating four Dutch word vector models. *Br. J. Educ. Technol.* 55, 1354–1375. doi:10.1111/bjet.13431

Gao, H., and Zhong, B. (2022). A blockchain-based framework for supporting BIM-based building code compliance checking workflow. *IOP Conf. Ser. Mater. Sci. Eng.* 1218, 012016. doi:10.1088/1757-899X/1218/1/012016

Guo, D., Onstein, E., and La Rosa, A. (2021). A semantic approach for automated rule compliance checking in construction industry. *IEEE Access* 9, 129648–129660. doi:10.1109/ACCESS.2021.3108226

Ismail, A., Ali, K., and Iahad, N. (2017). "A Review on BIM-based automated code compliance checking system," in 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), 16-17 July 2017, Langkawi, Malaysia. 1–6. doi:10.1109/ICRIIS.2017.8002486

Jadala, V., and Burugari, V. (2024). Artificial intelligence techniques in semantic web services. *Int. J. Innovation Multidiscip. Sci. Res.* 02, 42–49. doi:10.61239/ijimsr.2024.2220

Jiang, L., Shi, J., and Wang, C. (2022). Multi-ontology fusion and rule development to facilitate automated code compliance checking using BIM and rule-based reasoning. *Adv. Eng. Inf.* 51, 101449. doi:10.1016/j.aei.2021.101449

Kasim, T. (2015). *BIM-based smart compliance checking to enhance environmental sustainability*. PhD Thesis (Cardiff University).

Ke, J., Zacouris, Z., and Acosta, M. (2024). Efficient validation of SHACL shapes with reasoning. *Proc. VLDB Endow.* 17, 3589–3601. doi:10.14778/3681954.3682023

Li, S., Wang, J., and Xu, Z. (2024). Automated compliance checking for BIM models based on Chinese-NLP and knowledge graph: an integrative conceptual framework. *Eng. Constr. Archit. Manag.* doi:10.1108/ecam-10-2023-1037

Liu, X., Chen, Q., Wu, X., Hua, Y., Chen, J., Li, D., et al. (2020). Gated semantic difference based sentence semantic equivalence identification. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 28, 2770–2780. doi:10.1109/TASLP.2020.3030493

Mandal, S., Gandhi, R., and Siy, H. (2015). "Semantic web representations for reasoning about applicability and satisfiability of federal regulations for information security," in 2015 IEEE Eighth International Workshop on Requirements Engineering and Law (RELAW), 25-25 August 2015, Ottawa, ON, Canada. 1–9. doi:10.1109/RELAW.2015.7330205

Martinez-Gil, J., and Chaves-González, J. (2022). Sustainable semantic similarity assessment. *J. Intell. Fuzzy Syst.* 43, 6163–6174. doi:10.3233/jifs-220137

Motta, J., and Ladouceur, J. (2017). A CRF machine learning model reinforced by ontological knowledge for document summarization.

Nama, E., and Alalawi, A. (2023). "The adoption of automated building code compliance checking systems in the architecture, engineering, and construction industry," in 2023 International Conference On Cyber Management And Engineering (CyMaEn), 26-27 January 2023, Bangkok, Thailand. 289–296. doi:10.1109/CyMaEn57228.2023.10050930

Plansangket, S., and Gan, J. (2015). A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search. *Artif. Intell. Res.* 4, 119–125. doi:10.5430/AIR.V4N2P119

Poniszewska-Marańda, A., and Czechowska, E. (2021). Kubernetes cluster for automating software production environment. *Sensors (Basel, Switz).* 21, 1910. doi:10.3390/s21051910

Preidel, C., and Borrmann, A. (2018). BIM-based code compliance checking. In *Building Information Modeling*. 367–381. doi:10.1007/978-3-319-92862-3_22

Salama, D., and El-Gohary, N. (2011). Semantic modeling for automated compliance checking. *Computing in Civil Engineering*. 641–648. doi:10.1061/41182(416)79

Salama, D., and El-Gohary, N. (2013). Automated compliance checking of construction operation plans using a deontology for the construction domain. *J. Comput. Civ. Eng.* 27, 681–698. doi:10.1061/(ASCE)CP.1943-5487.0000298

Subramanian, V., Prince, G., Venkateshwara, S., Thangakumar, J. S., and Preetha, M. (2024). "Similarities and ranking of documents using TF-IDF, LDA and WAM," in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 18-19 April 2024, Chennai, India. 01–07. doi:10.1109/ADICS58448.2024.10533526

Sydora, C., and Stroulia, E. (2020). Rule-based compliance checking and generative design for building interiors using BIM. *Automation Constr.* 120, 103368. doi:10.1016/j.autcon.2020.103368

Tang, K., Fei-Fei, L., and Koller, D. (2012). "Learning latent temporal structure for complex event detection," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 16-21 June 2012, Providence, RI, USA. 1250–1257. doi:10.1109/CVPR.2012.6247808

Tierney, P. J. (2012). A qualitative analysis framework using natural language processing and graph theory. *Int. Rev. Res. Open Distributed Learn.* 13, 173–189. doi:10.19173/IRRODL.V13I5.1240

Westin, F. (2024). Time Period categorization in fiction: a comparative analysis of machine learning techniques. *Cataloging & Classif. Q.* 62, 124–153. doi:10.1080/01639374.2024.2315548

Xue, X., and Zhang, J. (2022). Regulatory information transformation ruleset expansion to support automated building code compliance checking. *Automation Constr.* 138, 104230. doi:10.1016/j.autcon.2022.104230

Zakon.kz (2025). Online legislative database. Available online at: https://online.zakon.kz/Document/?doc_id=39721077&pos=1;-16#pos=1 February 11, 2025).

Zhang, J., and El-Gohary, N. (2013). Information transformation and automated reasoning for automated compliance checking in construction. *Computing in Civil Engineering*. 701–708. doi:10.1061/9780784413029.088

Zhang, J., and El-Gohary, N. (2015). Automated information transformation for automated regulatory compliance checking in construction. *J. Comput. Civ. Eng.* 29. doi:10.1061/(ASCE)CP.1943-5487.0000427

Zhang, J., and El-Gohary, N. (2016). Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. *J. Comput. Civ. Eng.* 30. doi:10.1061/(ASCE)CP.1943-5487.0000346

Zhang, J., and El-Gohary, N. (2017). Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. *Automation Constr.* 73, 45–57. doi:10.1016/J.AUTCON.2016.08.027

Zhang, J., and El-Gohary, N. (2017). Semantic-based logic representation and reasoning for automated regulatory compliance checking. *J. Comput. Civ. Eng.* 31. doi:10.1061/(ASCE)CP.1943-5487.0000583

Zhang, R., and El-Gohary, N. (2021). A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. *Automation Constr.* 132, 103834. doi:10.1016/j.autcon.2021.103834

Zhang, Z., Ling, M., and Tim, B. (2022). "Towards fully-automated code compliance checking of building regulations: challenges for rule interpretation and representation," in Proceedings of the 2022 European Conference on Computing in Construction. doi:10.35490/ec3.2022.148

Zheng, Z., Zhou, Y., Lu, X., and Lin, J. (2022). Knowledge-informed semantic alignment and rule interpretation for automated compliance checking. *Automation Constr.* 142, 104524. doi:10.1016/j.autcon.2022.104524

Zhong, B., Luo, H., Hu, Y., and Sun, J. (2012). Ontology-based approach for automated quality compliance checking against regulation in metro construction project. 385–396. doi:10.1007/978-3-642-27963-8_35

Zhong, B. T., Ding, L. Y., Luo, H. B., Zhou, Y., Hu, Y. Z., and Hu, H. M. (2012). Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking. *Autom. Constr.* 28 (2012), 58–70. doi:10.1016/j.autcon.2012.06.006