A Study on the Semantic Similarity Evaluation Method of Nuclear Facility Safety Regulation using natural language processing

Ieck-Chae Euom [a*], Joon-Seok Kim [b]
[a]*Department of Data Science, Chonnam National University, Gwangju 61186, Republic of Korea*
[b]*System Security Research Center, Chonnam National University, Gwangju 61186, Republic of Korea*
[*]*Corresponding author : icelaken@ chonnam.ac.kr*

***Keywords : Natural Language Processing (NLP), Regulatory Analysis, Safety Requirements**

1. Introduction

The safety and security of nuclear facilities have historically developed as separate domains, with safety focusing primarily on managing physical hazards and security on preventing unauthorized access and attacks. However, with the advent of digital systems, there is a growing need to consider cybersecurity and safety in an integrated manner. Cyberattacks on digital systems within nuclear facilities can escalate beyond mere security breaches, posing significant risks to the overall safety of these critical infrastructures. Therefore, an integrated approach that simultaneously addresses cybersecurity and safety is essential.

Traditionally, safety regulations for nuclear facilities have emphasized physical security, but the increasing prominence of cybersecurity has led international bodies like IEEE to strengthen regulations that encompass both aspects. These regulations are now recognized as crucial for maintaining the operational integrity of nuclear facilities, ensuring that both safety and cybersecurity are managed in a unified framework.

However, challenges remain in systematically linking cybersecurity with safety regulations. Specifically, the relationship between South Korea's RS-015 cybersecurity standard and international IEEE safety regulations has not been clearly defined, and there is a lack of comprehensive strategies to apply these connections effectively.

In this paper, we aim to perform a categorization process of safety requirements as a preliminary step before applying cybersecurity measures based on safety design. Through this approach, we seek to ensure that safety and cybersecurity are effectively integrated within nuclear facilities.

2. Background

2.1. Natural Language Processing (NLP)

Natural Language Processing is a technology that enables computers to understand and analyze human language. NLP processes text or speech data, extracting meaning or analyzing grammatical structures. This allows computers to classify text, translate, summarize, or perform sentiment analysis. NLP is a subfield of artificial intelligence (AI) that utilizes machine learning and deep learning to comprehend context and recognize complex language patterns. In this study, NLP is used to analyze security and safety regulations in PDF format.

2.2.1. Related Work

Elluri et al. utilized natural language processing (NLP) techniques to extract and compare textual similarities between regulatory documents and privacy policies. Their framework employs NLP to measure the semantic similarity between the General Data Protection Regulation (GDPR) and various cloud privacy policies. The extracted information is stored in a knowledge graph, offering a solution to automate the compliance process.[1]

Baekgyu Kwon et al. applied NLP to analyze unstructured design guidelines in the manufacturing industry. This approach focuses on extracting relevant information from documents and systematically organizing it into a knowledge base. This method supports the automated updating of design requirements, ensuring that guidelines are accurately followed.[2]

Kotal et al. combined NLP with deep learning to evaluate the alignment between privacy policies and the NIST Cybersecurity Framework. Specifically, NLP is used to process and analyze text data, while deep learning models like BERT are employed to assess the semantic similarity between IoT device privacy policies and NIST requirements.[3]

Table 1. Analysis of related work

Ref.	Considering Multiple Regulations	Multimodal Approach Utilization	Requirement Relationship Analysis
[1]	O	X	O
[2]	X	O	X
[3]	O	X	O

Through this, my paper overcomes the limitations of previous research by conducting a comprehensive safety and security requirement analysis utilizing natural language processing (NLP). This research considers multiple regulations, incorporates a multimodal approach, and performs an in-depth requirement relationship analysis, addressing the gaps identified in previous studies.

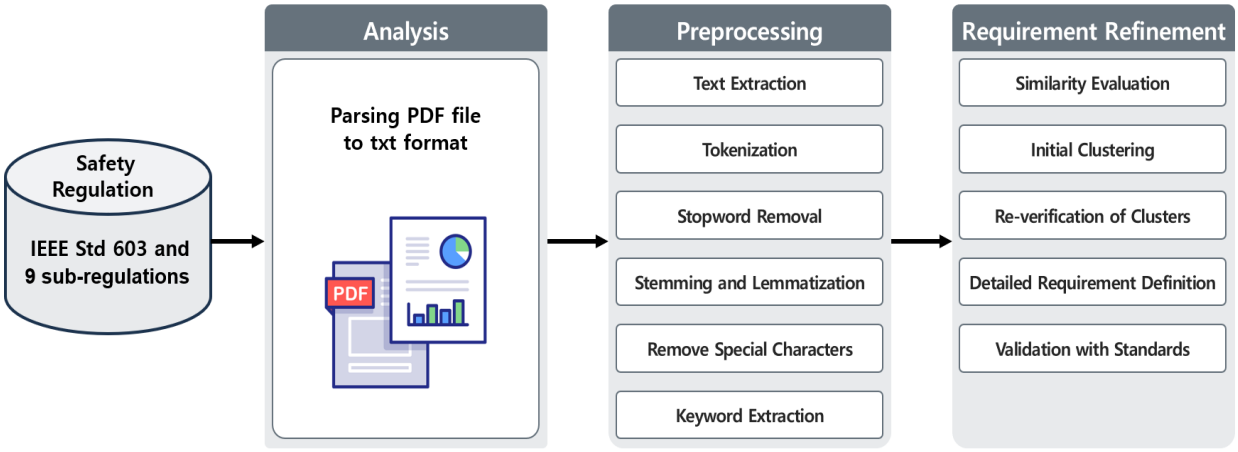

Figure 2. Categorization Methodology

3. Integrated Regulatory Analysis Method

In this paper, standards related to the safety system of nuclear facilities, including IEEE Std 603 and 9 additional standards, are selected. Design criteria and requirements are extracted from the main text, excluding covers and appendices.[4] The extracted items are verified for similarity using SBERT and TF-IDF, followed by initial categorization through Agglomerative Clustering. Subsequently, the similarity of the initially categorized items is re-verified to define the requirement names and details. Finally, the requirements are redefined using Bart and Attention Mask, based on key terms and sentences identified. Figure 1 illustrates the entire process.

3.1. Safety Regulation Analysis

Phase extracts and preprocesses requirements, then clusters them based on similarity.

3.1.1. Data Extraction and Preprocessing

This stage involves extracting and preprocessing key design criteria and requirements from nuclear facility safety standards (e.g., IEEE Std 603 and nine related sub-standards). PyPDF2 is used to extract text from PDF documents, and Tesseract OCR is employed to convert text from image-based documents. This process excludes unnecessary sections like covers, definitions, references, and appendices, focusing only on extracting meaningful content from the main text.

Once the text data is extracted, it undergoes preprocessing before being applied to natural language processing (NLP) techniques. The preprocessing involves the following steps:

Tokenization divides the extracted text into smaller units, such as words or sentences, using the nltk.word_tokenize function. This process is essential as it converts the text into a format that can be processed more efficiently in subsequent tasks.

Stopword Removal eliminates unnecessary words, often referred to as stopwords, using the nltk.corpus.stopwords library. By filtering out words like "the," "is," and "and," this step ensures that the focus remains on the essential content of the text.

Lemmatization converts words to their base forms with the help of nltk.WordNetLemmatizer. For example, "running" is transformed into "run," ensuring consistency across different variations of the same word.

Stemming extracts the root form of words, thereby reducing different variations of a word to a common base. This technique helps in reducing data size and improving the efficiency of text analysis.

Remove Special Characters is a process that eliminates special characters, symbols, and numbers from the text using regular expressions (re.sub). This step results in clean and analyzable data, which is crucial for further analysis.

The preprocessed text data is now ready for the next stage of similarity evaluation and clustering.

First, the selection of safety standards related to nuclear facilities was conducted. This included IEEE Std 603 and nine additional sub-standards, which are essential for regulating the safety of nuclear power plants. These documents were chosen as the basis for further analysis and requirement extraction.

Next, the relevant pages for analysis were identified within each standard document. During this process, non-essential sections such as covers, definitions, references, and appendices were excluded. The focus was placed on the main text containing the design criteria and requirements. The selected pages represent the key sections of each document that contain critical safety information.

Finally, design criteria and requirement items were extracted from the analyzed pages. Each standard document's specified requirements and regulatory criteria were categorized and counted as individual items. This data was then organized into a table, listing the standard name, analyzed page range, and the number of extracted items. This table clearly illustrates the number of requirements extracted from each document and the specific pages that were analyzed.

Table 3. Safety Requirements Analysis Scope

Standard		Analyzed Pages	Number of Extracted Items
IEEE Std 308	IEEE Standard Criteria for Class 1E Power Systems for Nuclear Power Generating Stations	p.23-p.58	31
IEEE Std 338	IEEE Standard Criteria for Periodic Surveillance Testing of Nuclear Power Generating Station Safety Systems	p.9-p.16	12
IEEE Std 379	IEEE Standard for Application of the Single-Failure Criterion to Nuclear Power Generating Station Safety Systems	p.17-p.23	10
IEEE Std 384	IEEE Standard Criteria for Independence of Class 1E Equipment and Circuits	p.14-p.33	34
IEEE Std 420	IEEE Standard for the Design and Qualification of Class 1E Control Boards, Panels, and Racks Used in Nuclear Power Generating Stations	p.15-p.22	14
IEEE Std 497	IEEE Standard Criteria for Accident Monitoring Instrumentation for Nuclear Power Generating Stations	p.20-p.30	34
IEEE Std 577	IEEE Standard Requirements for Reliability Analysis in the Design and Operation of Safety Systems for Nuclear Power Generating Stations	p.14-p.18	8
IEEE Std 603	IEEE Standard Criteria for Safety Systems for Nuclear Power Generating Stations	p.19-p.31	43
IEEE Std 741	IEEE Standard for Criteria for the Protection of Class 1E Power Systems and Equipment in Nuclear Power Generating Stations	p.23-p.58	9
IEEE Std 7-4.3.2	IEEE Standard Criteria for Programmable Digital Devices in Safety Systems of Nuclear Power Generating Stations	p.14-p.44	64
Total			259

3.1.2. Similarity Evaluation and Categorization

Building on the preprocessed text data, the next step involves evaluating the similarity between safety requirements. This stage leverages advanced natural language processing (NLP) techniques to analyze the semantic similarity between requirements. Sentence-BERT is utilized to convert requirement sentences into vectors, allowing the similarity between sentences to be measured in vector space. This method effectively captures the complex semantic relationships between sentences.

Additionally, the TF-IDF (Term Frequency-Inverse Document Frequency) technique is employed to calculate the importance of words within each document. By evaluating the frequency of words and their inverse document frequency, this method assigns weights to words, enabling a more precise measurement of similarity between documents. Cosine Similarity is then used to measure the similarity between document vectors represented by TF-IDF.

Once the similarity evaluation is complete, Agglomerative Clustering is applied to group similar requirements. This clustering method systematically groups the most similar requirements together based on their evaluated similarity, forming clusters that reflect the relationships between different requirements. This process allows for a more structured and organized representation of the requirements within the regulatory documents.

After clustering, the consistency of each cluster is reviewed, and additional clustering may be conducted if necessary. During this stage, requirement names and details are defined, ensuring that the various requirements within the safety regulations are clearly and consistently restructured.

The word cloud is generated by calculating the frequency of each word in the preprocessed text and using the WordCloud library. The size of each word in the cloud is determined by its frequency, with more frequent words appearing larger. This visual representation emphasizes the most important words in the document, allowing for quick identification of key themes. Figure 2 illustrates the word cloud.

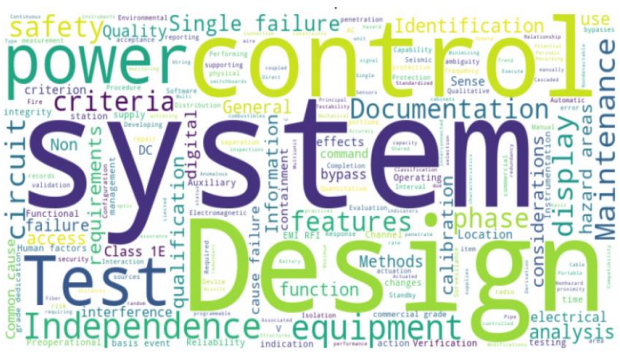

Figure 2. Word Cloud of Requirements

Based on the preprocessed text data, the frequency of each word is calculated. This step involves determining how many times each word appears in the text, providing insight into the most commonly used words in the document. This process utilizes the CountVectorizer from Python's Scikit-learn library. CountVectorizer processes the text data, calculates the frequency of each word, and converts it into numerical data. The word frequencies are then sorted in descending order, and this information is used to create a frequency table. Table 2 shows the frequency of each word.

Table 4. Safety Requirements Document Key Words Frequency Analysis

Word	Frequency	Word	Frequency	Word	Frequency
shall	1075	may	152	associated	96
safety	498	failure	144	devices	93
system	471	software	143	accident	93
equipment	331	testing	139	plant	89
design	310	analysis	136	supply	89
1e	292	used	133	operation	89
power	275	provided	129	circuit	86
systems	265	functions	129	within	85
class	244	redundant	118	protective	84
requirements	240	following	117	time	81
control	211	type	112	data	81
std	204	basis	112	single	81
ieee	204	include	107	acceptable	80
required	194	protection	106	conditions	80
function	194	separation	104	use	79
circuits	185	failures	103	cables	74
test	162	features	97	criteria	74

3.2. Requirement Refinement and Definition

This stage involves the re-evaluation and final definition of requirements that were initially extracted and clustered in phase 3.1. The goal is to ensure that the requirements are clearly defined, consistent, and practically applicable within the regulatory framework. The process refines the requirements to align with the overall objectives of the safety standards, ensuring they are ready for implementation.

3.2.1. Re-verification of Categorized Requirements

In this step, SBERT (Sentence-BERT), an advanced natural language processing (NLP) technique, is used to re-evaluate the initially clustered requirements from phase 3.1. SBERT excels at evaluating semantic similarities between sentences, which allows for a thorough review of the relationships between requirements within each cluster. This process identifies and corrects any inconsistencies, overlaps, or potential conflicts among the requirements. If necessary, additional re-clustering is performed to ensure that the requirements are logically and coherently organized.

This step ensures that the clustered requirements are more accurately categorized, reducing the likelihood of redundancy or conflicts. Each cluster is refined to focus on a clear and consistent theme, ensuring that the requirements are meaningful and can be applied without confusion within the regulatory framework.

3.2.2. Requirement Redefinition

Building on the re-verified requirements, this step focuses on the final definition and refinement of the requirements. Latent Dirichlet Allocation (LDA), a topic modeling technique, is employed to identify the key themes within each cluster, providing a basis for further refinement. LDA is particularly useful for uncovering the underlying topics that group the clustered requirements, ensuring they are centered around specific, relevant themes. Additionally, Bart and Attention Mask models are utilized to refine and clarify the requirement titles and detailed descriptions. The Bart model is highly effective in text generation and summarization tasks, enabling a more concise and precise redefinition of the requirements.

The redefined requirements will be clear, consistent, and structured in a way that they can be practically implemented in safety systems. Each requirement will be aligned with the overall context of the regulatory document, ensuring practical applicability. Furthermore, these redefined requirements will be cross-referenced with the original standards to verify their accuracy and ensure they align with the foundational principles of the regulatory framework.

Table 5. Categorizing Safety System Requirements

Classification	Safety Control
Testing and calibration	Safety function integrity
	Configuration management
	Setpoint verification
	Periodic testing
Design of equipment and circuits	Minimization of negative impacts
	Single failure criterion
	Common cause failure criterion
	Physical and logical independence

	Minimization of complexity
Verification and validation (V&V)	Reliability and objectivity
	Testing timing
	Verification of commercial off-the-shelf (COTS) products
Configuration changes	Compliance with quality standards
	Assurance of reliability
Data communicationl	Separation of functions and domains
	Minimization of external impacts
	Integrity and reliability verification
	Control of data flow
	Boundary protection system
	Maintenance operations
Resource constraints	Execution of control logic
	Reception of control signals
	Performance limitations
Risk management	Risk identification and assessment
	Risk mitigation
Incident and emergency response	Provision of backup systems
	Monitoring system
	Fault detection and recovery functionality
Access control	Physical access control
	Cybersecurity control
	Access rights management

4. Conclusion

In conclusion, this study successfully performed a systematic categorization of safety requirements as a preliminary step before applying cybersecurity measures based on pre-design principles in nuclear facilities. This process has enabled clear definition and consistent organization of safety requirements, establishing a solid foundation for future cybersecurity applications.

For future research, we plan to analyze the correlation between these categorized safety requirements and the cybersecurity guidelines outlined in RG 5.71. This analysis will further strengthen the integrated approach to safety and security, and propose more structured and effective methods for implementing cybersecurity in nuclear facilities.

ACKNOWLEDGMENTS

'This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.IITP-RS-2022-II221203, Regional strategic Industry convergence security core talent training business, 50%) and Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No.2106061, 50%)'

REFERENCES

[1] Elluri, Lavanya, Karuna Pande Joshi, and Anantaa Kotal. "Measuring semantic similarity across eu gdpr regulation and cloud privacy policies." 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020..

[2] KWON, Baekgyu, et al. Construction of design requirements knowledgebase from unstructured design guidelines using natural language processing. Computers in Industry, 2024, 159: 104100.

[3] CHAUDHARY, Namrata. Evaluating the alignment of privacy policies to NIST cybersecurity framework using Natural Language Processing and Deep Learning. 2021.

[4] IEEE Std 603-2009 - IEEE Standard Criteria for Safety Systems for Nuclear Power Generating Stations