Towards a Framework for Identifying Relevant Information in regard to Specific Context on the Use Case of Standards and Directives

Max Layer[1,*], Sebastian Neubert[1], Brij Boda[1], Ralph Stelzer[2]

[1] Siemens Energy Global GmbH & Co.KG
[2] Virtual Product Development, Technical University Dresden

** Corresponding Author:*
Max Layer
Siemens Energy Global GmbH & Co.KG
Siemenspromenade 9
91058 Erlangen
☎ *0172/3483436*
✉ *max.layer@siemens-energy.com*

Abstract

In a complex landscape of engineering requirements and knowledge represented by standards and directives, navigating and interpreting context-specific information remains a substantial challenge. A novel approach to address this issue is presented, introducing an information extraction framework for identifying product and context-specific, relevant information within unstructured text. The methodology employs Natural Language Processing techniques in a pipeline to parse text from various resources. Different stages for introducing context are proposed based on a trade-off between speed, accuracy, and storage capacity. An initial test focuses on the identification of inspection requirements of piping, while illustrating other potential applications such as an external reference cluster.

Keywords

Text analysis, Standards, Information Retrieval, Clustering, Natural Language Processing

| DOI: 10.35199/dfx2023.24

1. Introduction and Motivation

The interpretation of technical knowledge embodied in documents, such as standards and directives, forms a fundamental principle for the design, practices, and procedures of product development across various industries. Far from being mere recommendations, these standards are often enforced by customers who require compliance in their tenders and specifications. Fulfillment of these standards is sometimes certified by independent licensing or regulatory bodies [1]. Official entities such as the European Union further mandate directives [2], which are required for product certification within a given jurisdiction, regardless of customer requirements. The complexity of extracting the required information is compounded when these directives refer to a range of standards and norms. In the engineering sector in particular, directives such as the Machinery Directive or the Pressure Equipment Directive form the basis for compliance and certification. The growth of computerized data, text and document analysis tools, particularly in recent years, has provided opportunities to streamline and optimize this time intensive process.

However, several hurdles inhibit companies from leveraging these advancements. Among these is the issue of possible corporate intellectual property (IP) drain. In knowledge-driven economies, information forms the basis of competitive advantage, leading businesses to exercise caution while sharing their internal information to mitigate potential misuse or unintended exposure. Recent advancements in using large language models for text analyzation in corporate environments come with the significant risk of IP drain, as these models, based on reinforcement learning from human feedback (RLHF) [3, 4], can inadvertently learn and disseminate proprietary information while interacting with the user.

While the use of locally installed, specialized data analysis tools seem to offer support, these tools often require specific data formats for effective processing, leading to additional time and resources spent on data cleaning and preprocessing. Moreover, they come with a steep learning curve due to their multitude of features and functionalities, necessitating extensive training for staff members. This complexity can lead to delays and inefficiencies as organizations adapt to incorporate these tools into their existing workflows.

The final significant challenge lies in the context-specific nature of the information contained within technical documents. They are custom-made to suit a diverse array of products, processes, and practices, demanding a keen understanding of the product or situation's context for relevant information extraction. Any effort to extract relevant information necessitates a deep understanding of the specific context of the product or situation in question, embodied by an engineer or expert. The dynamic nature of regulatory environments further complicates this, as companies must keep pace with the constantly evolving norms and standards that mirror shifts in technology, societal expectations, and regulatory focus.

This paper attempts to propose a novel framework (see Figure 1), to mitigate these challenges. The objective is to provide a robust aid for interpreting unstructured technical knowledge from standards and directives by using product and specific context as an input. Furthermore, it intends to offer a solution that balances the need for fast information retrieval and adequately addresses the issues mentioned above.

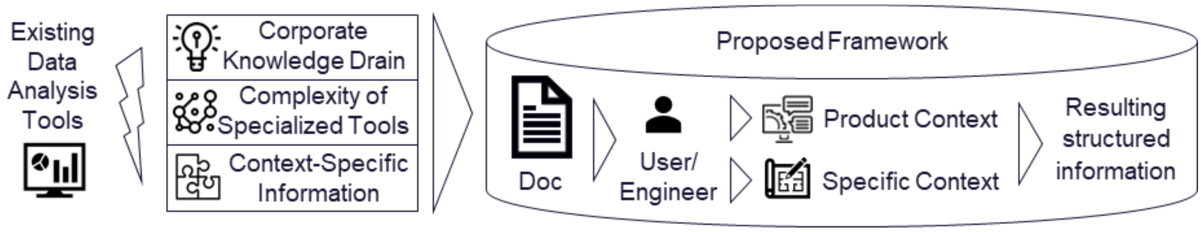

Existing Data Analysis Tools
Corporate Knowledge Drain
Complexity of Specialized Tools
Context-Specific Information
Proposed Framework
Doc
User/ Engineer
Product Context
Specific Context
Resulting structured information

Figure 1: Challenges and proposed framework for identifying relevant information from unstructured text

2. State of the Art

For the development of the concept, literature and patents were reviewed in detail and continuously validated with the possible application and feasibility in the corporate environment of process industry engineering. A special focus was put on existing information extraction approaches, natural language processing as well as information storage in general.

A similar research field covers the analysis of academia in regard to identifying relevant papers based on citation relations [5, 6]. Framed as Technical Language Processing (TLP) by Brundage et al. [7] focus on the application of Natural Language Processing (NLP) as a part of Machine Learning (ML) on technical text. Dunn et al. [8] presented a sequence-sequence approach to joint named entity recognition and relation extraction from complex information in scientific text. Anantharangachar et al. proposed an ontology-based information extraction which lists relevant domain ontologies created by human experts, extracts information in the form of triples and stores them into a Resource Description Framework [9]. In a similar approach, Xu, Qi et al. try to identify sensitive information within unstructured text by utilizing convolutional neural networks [10]. Lastly, Bourbakis and Mertoguno propose a holistic approach for analyzing technical documents with modality mapping [11]. Weston et al used NER for identifying inorganic material mentions, sample descriptors, phase labels, material properties and applications from materials science documents [12].

Besides existing approaches for the specific task of information retrieval, a few established tools will be elaborated in this chapter. Furthermore, methods for structured storage and evaluation metrics will be introduced.

2.1. Information Analysis Tools and Approaches

As the field advances, the development of more sophisticated tools and approaches is anticipated. With the proliferation of unstructured data, especially text, several tools and approaches have been developed to facilitate information extraction, analysis, and understanding. Apart from BERT, which will be elaborated more in depth in this paper, Named Entity Recognition (NER), focusing on identifying and classifying named entities in text, and Topic modelling based on Latent Dirichlet Allocation (LDA) have been used. LDA presumes a bag-of-words representation of a document and assumes that each document is a mixture of a certain number of topics, while each topic is a distribution over the entire set of words in the corpus.

BERT (Bidirectional Encoder Representations from Transformers), developed by Google, represents a significant advancement in the field of NLP. BERT is based on the Transformer architecture, which uses self-attention mechanisms to understand the context of words in a sentence bidirectionally. [13] BERT has revolutionized many NLP tasks, including sentiment analysis [14], named entity recognition [15], and question-answering [16], due to its ability to comprehend the nuances of language and the context of word usage. Shown in Figure 2, BERT's transformative architecture leverages a bidirectional training process and multiple transformer blocks, imbued with a multi-head self-attention mechanism and a position-wise fully connected feed-forward network, to comprehensively understand the context and relevance of each word in a sentence. BERTs bidirectional understanding allows it to capture the context and meaning of words more effectively, leading to better language understanding and representation. Lastly, BERT, available in different sizes, uses pre-training on extensive text data and subsequent fine-tuning to create highly versatile, task-agnostic models, demonstrating unprecedented efficacy in diverse NLP tasks without necessitating architecture modifications. [13]

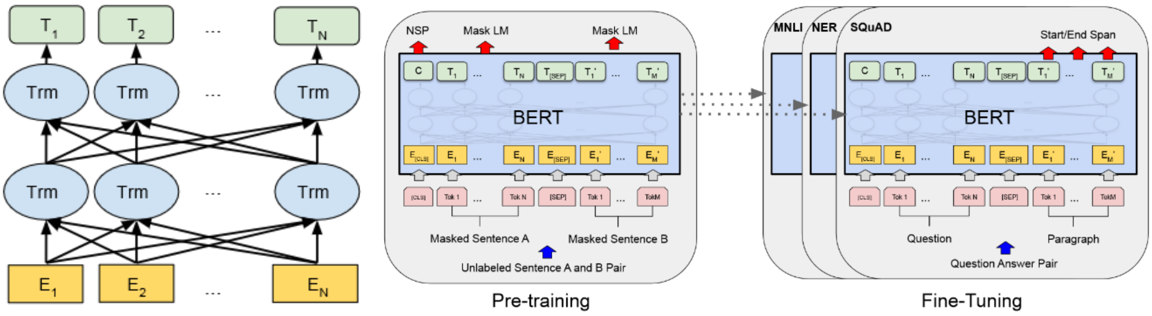


Figure 2: BERT Architecture as described by [13]

2.2. Methods for Structured Storage

Structured data storage is a crucial aspect of modern data systems, providing a means to organize, access, and utilize vast quantities of data efficiently. One of the aspects of storing analyzed documents and their information would be to avoid computationally expensive text extraction and analysis step. Relational Databases have long been the cornerstone of structured data storage. They store data in tables, where each row represents a record, and each column signifies an attribute. SQL (Structured Query Language) is the standard language used for interacting with these databases. MySQL, PostgreSQL, and Oracle Database are common examples of relational databases. Their robustness, reliability, and wide range of applications have made them a mainstay in data storage.

2.3. Evaluation Metrics for Machine Learning in NLP Context

NLP tasks such as document classification, NER and contextual information identification in general require specialized evaluation metrics. This chapter discusses the state-of-the-art evaluation metrics specifically relevant to these types of tasks. Metrics from both the classification and ranking domains could be relevant here, including accuracy, precision, recall, F1-score, AUC-ROC for classification and precision@k, recall@k, and mean average precision for ranking. [17]

Accuracy measures the proportion of correct predictions in total predictions. While simple, it may be misleading in imbalanced datasets. Precision is calculating the proportion of true positives among all positive predictions while Recall is calculating the proportion of true positives among actual positive instances, essential when false positives and false negatives carry different costs. The F1-score is the harmonic mean of precision and recall, balancing the two measures. The AUC-ROC metric is particularly useful in binary classification problems. It measures the classifier's ability to distinguish between classes and works well with imbalanced datasets.

3. Research Objective

From an engineering perspective, e.g., developing complex multidisciplinary products, information about requirements depends both on the product and the specific context. Interpreting and navigating through a landscape of standards and directives, especially when trying to identify information relevant to this specific context, can often prove to be time consuming and ineffective. In advance to proposing a novel approach to identifying context-specific relevant information within unstructured text, three research questions have been identified, shown in Figure 3. Furthermore, an approach for establishing and cross navigating through a reference cluster is included. With the proposed method, the authors provide the foundation for an information extraction framework that can be applied to different scenarios.

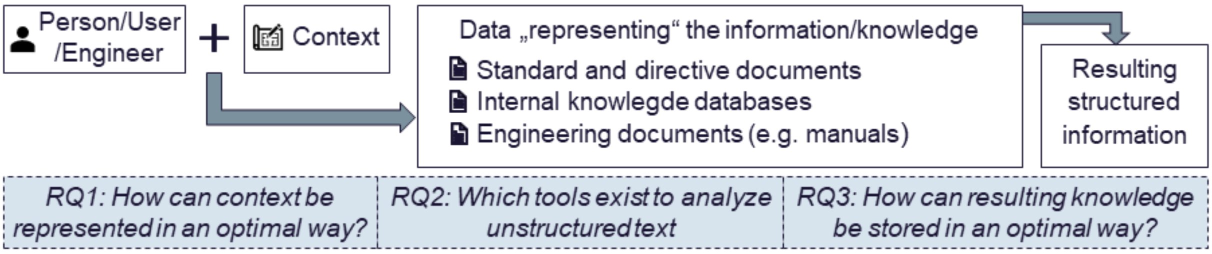

Person/User /Engineer
+
Context
Data „representing" the information/knowledge
Standard and directive documents
Internal knowlegde databases
Engineering documents (e.g. manuals)
Resulting structured information
RQ1: How can context be represented in an optimal way?
RQ2: Which tools exist to analyze unstructured text
RQ3: How can resulting knowledge be stored in an optimal way?

Figure 3: Interrelationships of individual, context and data with resulting research questions

4. Methodology

The python programming language [18] was used as the primary tool for implementing the proposed approach and conducting the experiments. Since the information is contained in semantic representation, NLP (see chapter 2.1) is utilized, which can help machines to process and understand human languages. Several NLP libraries support the development and implementation process, including the Natural Language Toolkit [19], spaCy [20], Scikit-Learn. For analyzing, grouping and extracting meaning of parsed text through above mentioned libraries, different techniques can be used such as sentence classification to differentiate between general information, technical information, general requirement and technical requirement sentences. Name Entity Recognition is utilized to identify different entities present in the sentences like component/systems, internal and external reference, person and processes mentioned. Topic modelling is also useful to identify underlying topic clusters within a text which can be further used to identify interrelationships between them.

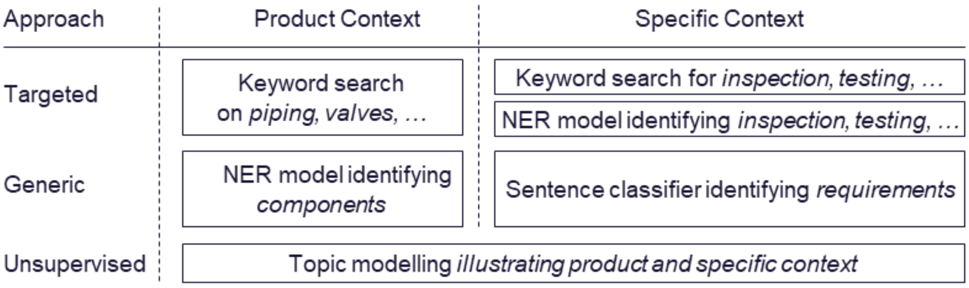
4.1. Preprocessing

Text extraction: Since most of the standards, norms, and directives are available in the form of PDF files, all proposed different pipelines start with text extraction from the pdfs. After reviewing common libraries that can be utilized, the library Pytesseract, based on [21], was able to perform highly clean text extraction jobs. As it works with optical character recognition (OCR), paragraphs are identified more clearly and it can also extract text from scanned documents, whereas other libraries may only be useful for text-based, searchable PDFs.

Cleaning and breaking text into sentences: To prepare the extracted text for performing different NLP tasks, it is necessary to divide it into sentences and sometimes to tokens. Here we utilized Fuzzy matching and Regular expressions (Regex) [22] for pattern matching task for cleaning unwanted patterns in the text such as headers, footers, unnecessary spacings. It will provide clean sentences as output.

4.2. Representation of Context

As briefly mentioned in section 1, the authors propose to divide the context into product context and specific context. The product context is introduced either by specifically defining component or systems names based on a Bill of Material (BoM) or by a NER model trained and able to identify components in general. These can then be filtered according to the needs of the engineer/user. The specific context on the other hand depends on the use case of the user/engineer. As described, the goal of the framework is to offer a solution for a targeted search, a generic overview or even an unsupervised understanding of the contents of a document, visualized in Figure 4. The exact retrieval of information is provided by boolean information retrieval, a binary-based paradigm that effectively filters pertinent materials based on exact match queries but has limited ranking capabilities. While vector space and probability models are more advanced but may have drawbacks when dealing with huge datasets or fluctuating word frequencies, boolean information retrieval is effective and simple for accurate matches.



Approach	Product Context	Specific Context
Targeted	Keyword search on *piping, valves, …*	Keyword search for *inspection, testing, …* NER model identifying *inspection, testing, …*
Generic	NER model identifying *components*	Sentence classifier identifying *requirements*
Unsupervised	Topic modelling *illustrating product and specific context*	

Figure 4: Different approaches for context representation and targeting.

4.3. Modular Pipeline consisting of Single Models

As described in the beginning of section 4, the main chosen modular elements consist of sentence classification, named entity recognition, intelligent keyword search and topic modelling. In the process of identifying relevant information, context can possibly be introduced at different stages of the information retrieval process, represented by analyzing different data pipelines (Figure 5). The first pipeline assumes that there is already specific product context and can therefore be filtered on list of components. Afterwards, the results are filtered through sentence classification further reducing. The second pipeline starts with sentence classification and can then be, based on modelled topics, filtered on components and specific context. The third pipeline is the most generic one and suggest an initial topic modelling, which can further be filtered, based on identified named entities.

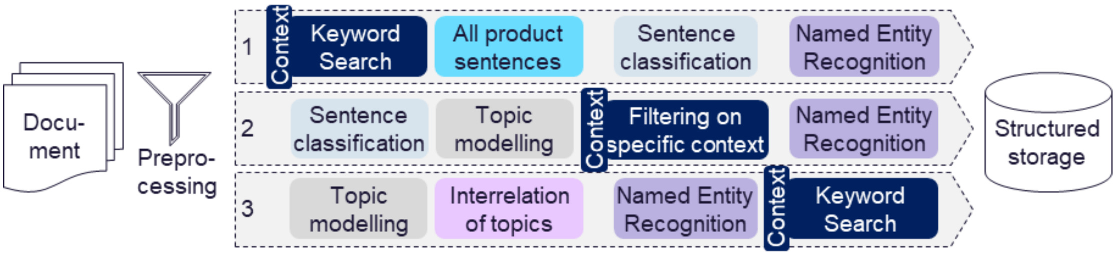
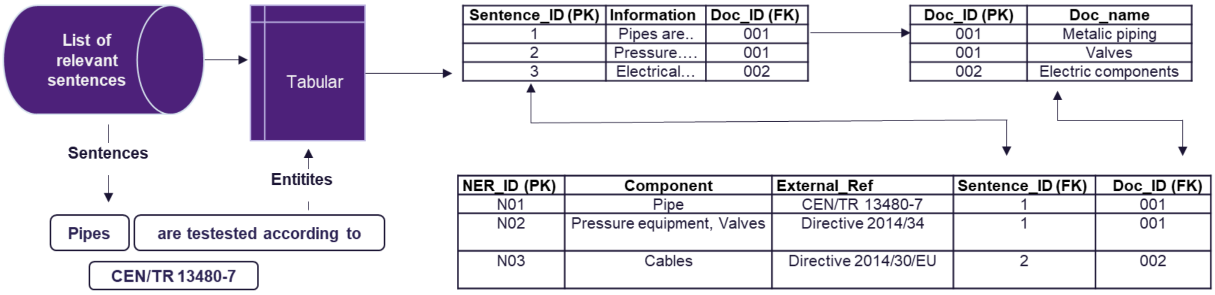


Figure 5: Pipeline approaches

4.4. Structured Display and Storage

In all cases, a structured database is created, which can then be used to retrieve the relevant information based on a specific context. As the framework is not intended to fully automate the process of information retrieval but rather efficiently shows a perspective user, where to look and how to proceed with the document analysis, the underlying storage is sentence centered. Based on user-generated searches, all the pertinent sentences that were located within a document are saved. The initial sentences will be saved in a tabular form (excel). The NER model [2] will be utilized in parallel to extract elements like components, tests, external references, people, and processes. With these methods, pertinent entities from the document are extracted and retrieved into the same excel sheet. Once these sheets are filled with enough documents, they can be uploaded to Postgre DBMS (Figure 6) for classification and centralized storage. By doing so, the processing step for users with similar requirements can be skipped.



Sentence_ID (PK)	Information	Doc_ID (FK)
1	Pipes are..	001
2	Pressure....	001
3	Electrical...	002

Doc_ID (PK)	Doc_name
001	Metalic piping
001	Valves
002	Electric components

NER_ID (PK)	Component	External_Ref	Sentence_ID (FK)	Doc_ID (FK)
N01	Pipe	CEN/TR 13480-7	1	001
N02	Pressure equipment, Valves	Directive 2014/34	1	001
N03	Cables	Directive 2014/30/EU	2	002

Figure 6: DBMS architecture for storing sentence centered.

4.5. Reference Cluster

To finalize the analysis of a given document, an additional focus is set on identifying internal and external references. As both identified reference and entities such as components or requirements are all linked to the sentence_id, a resulting topic cluster can be generated. To show the possibilies of a reference cluster, a dataset is exported from a standards database consisting of *norm_id*, *norm_name* and *referenced_id*. This was plotted and illustrated with the python library NetworkX [23] shown in Figure 7. For one analyzed document, based the sentence centered approach, the external references identified can be followed based on the tagged product and specific context in the proximity. This gives the user the possibility to identify which standards are relevant in regard to which product and specific context, enabling a direction for further research.

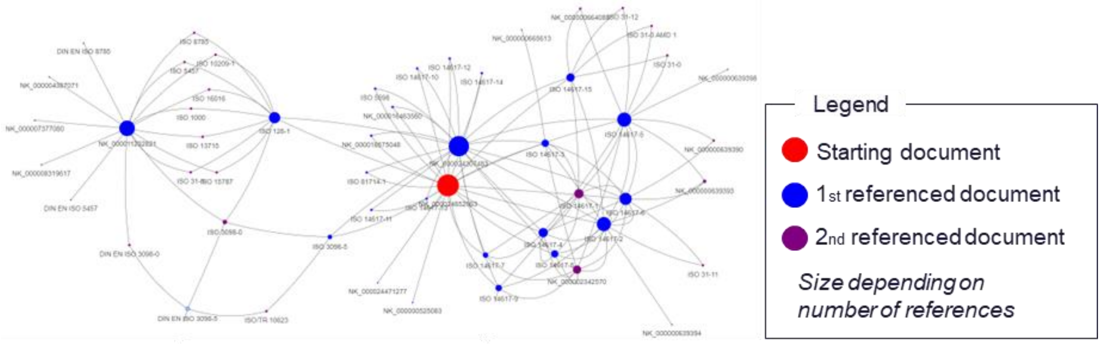


Figure 7: Reference cluster for external references

5. Results and Discussion

As an evaluation approach, different engineering standards and a directive are chosen to validate the concept. Furthermore, a discussion on the proposed storage options is given. The exemplary product context within this validation focuses on the component level on **piping**, **valves** and formed parts and on system level on **pressurized systems**. The specific context will generically investigate **requirements** stated in the documents further targeting **inspection, testing, maintenance** and **documentation**.

In the development of the framework, an approach was tested where context was provided to represent the product in the background by using the Bill of Material (BoM) itself. The process began with the transformation of the BoM in an XML-format into a structured format using a preprocess function. The dataset consists of three standards and one directive from process technology background:

- DIN EN 13480-5: Metallic industrial piping – Part 5: Inspection and testing [24]
- DIN EN 12569: Industrial valves – Valves for chemical and petrochemical process industry – Requirements and tests [25]

- EN ISO 10434 – Bolted bonnet steel gate valves for the petroleum, petrochemical and allied industries [26]
- Pressure directive EU 2014/68/EU [27]

5.1. BERT Sentence Classifier

To show the capabilities of a BERT model used for sentence classification, the use case of identifying defined requirements in text was evaluated. The model's goal is to determine whether or not a given statement constitutes a requirement. Contextual embeddings and attention processes in BERT allow it to recognize complex word associations, leading to precise and effective sentence classification performance. While this

The model was trained on a dataset of 208 manually labeled sentences with 4 epochs and then compared with the result data with a python script. The mean and standard deviation for all 4 documents can be found in Table 1. The results for the single model evaluation of the sentence classifier show sufficient accuracy for correctly identifying requirements within unstructured text.

Table 1: Evaluation of classified sentences

Parameters			**Results**				**Metrics**				
Document	**Sentences**	**Require-ments**	**TP**	**TN**	**FP**	**FN**	**Accuracy**	**Recall**	**Precision**	**F1**	**AUC-ROC**
DIN EN 13480-5	356	193	181	138	25	12	0,90	0,94	0,88	0,91	0,89
DIN EN 12569	254	145	142	95	14	3	0,93	0,98	0,91	0,94	0,93
DIN EN ISO 10434	543	251	215	261	31	36	0,88	0,86	0,87	0,87	0,88
2014_68_EU	1494	704	654	706	84	50	0,91	0,93	0,89	0,91	0,91
						Mean	**0,90**	**0,93**	**0,89**	**0,91**	**0,90**
						Standard deviation	**0,02**	**0,04**	**0,01**	**0,03**	**0,02**

5.2. NER Model

Named entity recognition (Figure 8) based on spaCy was utilized to identify system/components, process, person, location, test, internal and external references. This is used to structure the remaining sentences after classification and keyword search and for filtering. Boolean information retrieval model to find the exact matches through keyword search. The defined entities for the NER model are components, material, no+units, role, location, testing, date, process and internal and external references for clustering. An initial testing on exemplary identified entities is shown in Table 2.

Table 2: Evaluation of correctly identified occurrences with NER model

Overview			**Results**				**Metrics**			
Entity	**Sentences**	**Occurences**	**TP**	**TN**	**FP**	**FN**	**Accuracy**	**Recall**	**Precision**	**F1**
System/Components	356	562	522	9450	289	40	0.97	0.93	0,64	0.76
External Reference	356	438	431	9582	238	7	0,98	0,98	0,64	0,78

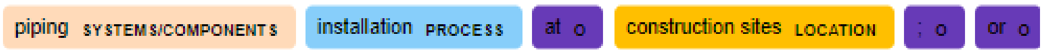


Figure 8: Exemplary visualization of NER predicted tagging.

5.3. Topic Modelling

The gensim topic modelling considers the given text as whole. After lemmatization, data to word conversion, corpus creation and LDA model set-up, the model can be visualized (Figure 9). The time for executing this pipeline depends on the computational power and the length of text. Based on a local CPU, different stages for introducing this pipeline were compared to evaluate the feasibility of using topic modelling at the beginning, middle or end of a document analysis process. The execution time for an already reduced document_1 to 196 requirement sentences took 5 minutes to run, while with 2 sentences it takes 5 seconds. Even though, this can be accelerated with more computing power, the authors recommend integrating topic modelling after preprocessing as a standard step before interacting with the document.

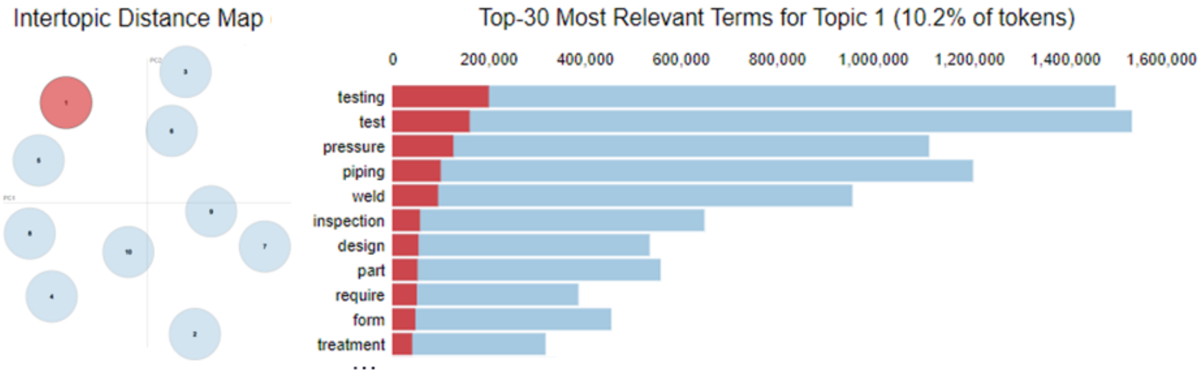


Figure 9: Plotted topic modelling based on DIN EN 13480-5

6. Summary and Outlook

While text analysis, interpretation and information retrieval are gaining popularity among different domains, these tools are not yet utilized to a sufficient amount within engineering and technical environments. Therefore, the foundations for a generic framework for identifying relevant information within unstructured text is proposed. The authors highlight the challenges posed by the complexity and context-specific nature of technical documents and the need for efficient information extraction. The framework utilizes Natural Language Processing to parse text and extract context-specific information. The authors suggest introducing context at various stages in the information retrieval process, depending on the use case. The initial test of the framework focuses on identifying different components, requirements and references, therefore demonstrating its potential industrial application. Additionally, an outlook on the possibility of creating reference clusters based on the approach is given. The framework can be expanded to cover other domains and use cases, offering a comprehensive solution for interpreting and applying unstructured technical knowledge.

References

[1] Nair S, La Vara JL de, Sabetzadeh M, Falessi D. Evidence management for compliance of critical systems with safety standards: A survey on the state of practice. Information and Software Technology 2015;60:1–15.

[2] European Union. European Union directives. Available at: https://eur-lex.europa.eu/EN/legal-content/summary/european-union-directives.html. Accessed: 5 Jul 2023.

[3] Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Christiano P, Irving G. Fine-Tuning Language Models from Human Preferences, 2019.

[4] Lambert N, Werra L von. Illustrating reinforcement learning from human feedback (RLHF). Hugging Face Blog 2022.

[5] Liang Y, Li Q, Qian T. Finding Relevant Papers Based on Citation Relations. In: Wang H, Li S, Oyama S, Hu X, Qian T, editors. Web-Age Information Management. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011:403–14.

[6] Totti LC, Mitra P, Ouzzani M, Zaki MJ. A Query-oriented Approach for Relevance in Citation Networks. In: Bourdeau J, Hendler JA, Nkambou RN, Horrocks I, Zhao BY, editors. Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion. New York, New York, USA: ACM Press, 2016:401–06.

[7] Brundage MP, Sexton T, Hodkiewicz M, Dima A, Lukens S. Technical language processing: Unlocking maintenance knowledge. Manufacturing Letters 2021;27:42–46.

[8] Dunn A, Dagdelen J, Walker N, Lee S, Rosen AS, Ceder G, Persson K, Jain A. Structured information extraction from complex scientific text with fine-tuned large language models, 2022.

[9] Anantharangachar R, Ramani S, Rajagopalan S. Ontology Guided Information Extraction from Unstructured Text. IJWesT 2013;4:19–36.

[10] Xu G, Qi C, Yu H, Xu S, Zhao C, Yuan J. Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network. In: 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC): IEEE, 2019:474–79.

[11] Bourbakis N, Mertoguno S. A Holistic Approach for Automatic Deep Understanding and Protection of Technical Documents. Int. J. Artif. Intell. Tools 2020;29:2050007.

[12] Weston L, Tshitoyan V, Dagdelen J, Kononova O, Trewartha A, Persson KA, Ceder G, Jain A. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. J Chem Inf Model 2019;59:3692–702.

[13] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: arXiv, 2018.

[14] Li X, Bing L, Zhang W, Lam W. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis, 2019.

[15] Hakala K, Pyysalo S. Biomedical Named Entity Recognition with Multilingual BERT. In: Jin-Dong K, Claire N, Robert B, Louise D, editors. Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019:56–61.

[16] Wang Z, Ng P, Ma X, Nallapati R, Xiang B. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering, 2019.

[17] Harrison M. Machine learning pocket reference: working with structured data in python: O'Reilly Media, 2019.

[18] van Rossum G, Drake FL. Python reference manual: Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[19] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit: " O'Reilly Media, Inc.", 2009.

[20] Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear 2017;7:411–20.

[21] Kay A. Tesseract: an open-source optical character recognition engine. Linux Journal 2007;2007:2.

[22] Aho AV. Algorithms for finding patterns in strings, Handbook of theoretical computer science (vol. A): algorithms and complexity. MA: MIT Press, Cambridge 1991.

[23] Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX.

[24] DIN EN 13480-5. Metallic industrial piping : Part 5: Inspection and testing;23.040.01(DIN EN 13480-5:2017). Switzerland: ISO copyright office, 2017. Accessed: 20 Jul 2023.

[25] DIN EN 12569. Industrial valves - Valves for chemical and petrochemical process industry: Requirements and tests;23.060.01(DIN EN 12569:2020). Switzerland: ISO copyright office, 2020. Accessed: 20 Jul 2023.

[26] DIN EN ISO 10434. Bolted bonnet steel gate valves for the petroleum, petrochemical and allied industries;75.180.20(DIN EN ISO 10434:2020). Switzerland: ISO copyright office, 2020. Accessed: 20 Jul 2023.

[27] 2014/68/EU. Pressure Equipment Directive(2014/68/EU): THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, 2014. Accessed: 15 Feb 2023.