Characterizing Regulatory Documents and Guidelines based on Text Mining

Karolin Winter[1], Stefanie Rinderle-Ma[1], Wilfried Grossmann[1], Ingo Feinerer[2], Zhendong Ma[3]

[1] Faculty of Computer Science, University of Vienna, Vienna, Austria
{karolin.winter, stefanie.rinderle-ma, wilfried.grossmann}@univie.ac.at
[2] University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria
ingo.feinerer@fhwn.ac.at
[3] Center for Digital Safety & Security, Austrian Institute of Technology, Vienna, Austria
zhendong.ma@ait.ac.at

Abstract. Implementing rules, constraints, and requirements contained in regulatory documents such as standards or guidelines constitutes a mandatory task for organizations and institutions across several domains. Due to the amount of domain-specific information and actions encoded in these documents, organizations often need to establish cooperations between several departments and consulting experts to guide managers and employees in eliciting compliance requirements. Providing computer-based guidance and support for this often costly and tedious compliance task is the aim of this paper. The presented methodology utilizes well-known text mining techniques and clustering algorithms to classify (families) of documents according to topics and to derive significant sentences which support users in understanding and implementing compliance-related documents. Applying the approach to collections of documents from the security and the medical domain demonstrates that text mining is a promising domain-independent mean to provide support to the understanding, extraction, and analysis of regulatory documents.

Keywords: Compliance, Regulatory documents, Requirements extraction, Text mining

1 Introduction

Eliciting and implementing requirements from textual sources, i.e., regulations such as Basel III [13], represents a major challenge for todays businesses and requires a significant effort in terms of time and cost (cf., e.g., [18]). For example, for a company the average cost (respectively duration) to implement the ISO 27001 standard is estimated between $6500 and $26000 (respectively between 6 and 12 months) [16]. It is often a manual and tedious process to interpret, adapt, and implement the clauses from standards and guidelines into appropriate technologies, processes, and actions, supported by consultants. Moreover, the cooperation between consultants and several departments is required in order to correctly interpret and adapt these regulations. Most employees and managers are experts in their field but dealing with these cumbersome documents is a challenge. The sheer breadth, intentional neutrality, and lack of actionable details are the main obstacles to understand and apply the information in a meaningful way. On the other hand, since many of these documents went through a very rigorous drafting and voting process, the use of terminologies and words is carefully thought aiming for maximal precision. This raises an interesting question i.e., *what if computers can be used to assist to understand, interpret, and implement regulatory documents and guidelines and to extract the salient features?* A comprehensive approach towards this question has not been provided yet [21].

In order to access this problem the paper aims at answering the following research questions:

RQ1 *How can standard text mining tools help to understand the topics and content of regulatory documents and guidelines?*
RQ2 *How to improve the results produced by these methods?*
RQ3 *Is it possible to extract sentences which are relevant for implementing such documents?*

For this purpose, a methodology is proposed that constitutes a first step towards covering RQ1, RQ2, and RQ3. One idea of this methodology is the fragmentation of (larger) documents into subdocuments in order to exploit their logical structure and to improve the results of text mining techniques. Another idea is the further analysis of the documents using clustering to group document fragments by topics. The proposed methodology is evaluated based on three case studies whereupon two of them are described in the paper and the third one is added as supplementary material[4]. The first case study features a selection of ISO 27000 standard documents. Contrary, for the second case study only one medical document was chosen to demonstrate the usage of the document fragmentation contained in the methodology. Therefore, it will be possible to outline the feasibility and applicability of the approach to a variety of domains and documents. Privacy documents are the subject of the third case study.

The paper is structured as follows. First, in Sect. 2, the methodology of the approach is presented. Afterwards, the document collections used for the case studies are described in Sect. 3.1 and 3.2. The evaluation of these case studies is issued in Sect. 4 while in Sect. 5 use cases and limitations are discussed. Related work is presented in Sect. 6 and the paper terminates with a conclusion and outline of future work in Sect. 7.

2 Methodology

The methodology presented in this section is depicted in Figure 1 and was implemented in *R*[5]. The starting point is the collection of selected documents which

[4] `http://www.wst.univie.ac.at/projects/sprint/index.php?t=tm`

[5] `https://www.R-project.org/`

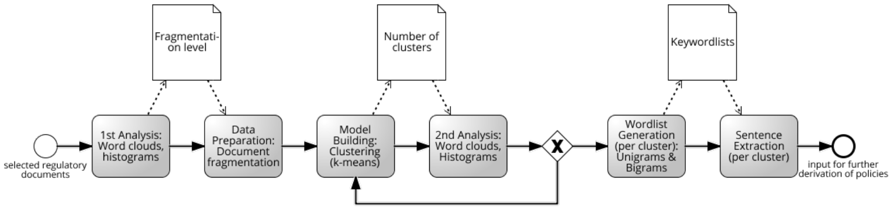


Fig. 1: Characterization of regulatory documents – Methodology

is loaded into one corpus. Standard corpus transformations as implemented in the *tm-package*[6] are applied on this basic corpus, for example converting all characters to lower case, removing the numbering, punctuation, stop words as well as customized (stop) words depending on the content and structure of the document (e.g., copyright labels).

To acquire an overview of the topics covered in this set of documents, frequent terms are resolved and represented by word clouds and histograms, applying on the one hand *weightTf* (a weighting scheme using term frequencies) and on the other hand *weightTfIdf* (term frequency — inverse document frequency) [23] (**1st Analysis**). Using *weightTfIdf* is especially suited for document sets of different length as frequencies are normalized [17]. In our context *weightTfIdf* is appropriate due to the wide length spread of the documents within the first case study (cf. Sect. 3.1; ↦ RQ1). Nevertheless, these methods will in general not lead to an in-depth understanding of each document or parts of documents thus further analysis is necessary.

In order to achieve improved analytical results, the idea is to divide the documents into logical units or fragments and to perform analysis based on clustered fragments (**Data Preparation**; ↦ RQ2). It is well-known that text mining techniques deliver better results when applied to a large number of short documents (with respect to the number of words), e.g., tweets, than on few but large documents [11] with a dense information value. Figure 2 provides a categorization of the selected document collections for the case studies (cf. Sect. 3) compared to tweets with respect to two parameters, i.e., number of words per document and number of documents (per collection). The intention is to design the fragmentation in such a way that the resulting partial documents imitate the characteristics of tweets, i.e., few words, but many single documents in order to thin out the information density of the original document. So the goal of the data preparation step is to fragment the documents in such a way that the outcome, i.e., all partial documents considered together is located in the upper left (grey) corner of the figure. Consequently, the fragmentation should be fine-granule enough, but without leading to overly short fragments.

Additionally, it seems to be meaningful to operate on "connected" fragments that reflect the structure of the document. Normally this is suggested by the

[6] `https://CRAN.R-project.org/package=tm`

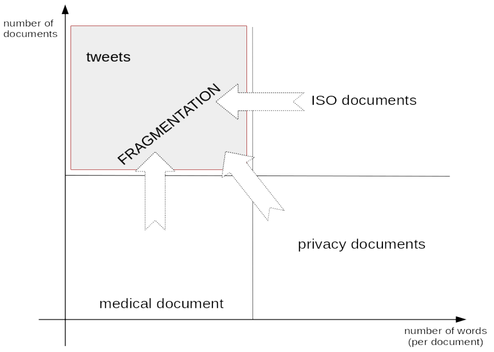


Fig. 2: Categorization of selected (collections of) documents based on their number and size, i.e., number of words per document

structuring of a text into sections and subsections. It depends on the documents which fragmentation level should be chosen. For implementation purposes it can be helpful if the documents have a clear structure and a table of contents, like the ISO documents (cf. R-Script in Sect. 4.1). In this case it is possible to extract (sub-) sections automatically based on the table of contents and suitable regular expressions.

After the fragmentation an optional further step is to merge all partial documents into one corpus on which the above mentioned transformations are applied once more. Now, frequent terms can be computed again in order to check if the splitting has led to a better and more detailed understanding of the documents' topics. Since as a next step clustering methods are used for resolving fragments treating similar topics (**Model building**), this could be helpful when a decision on the number of clusters has to be made.

Why is clustering of the fragments in this case feasible? Families of documents containing guidelines often have content overlaps so it is possible that fragments of different documents may treat the same topic and should therefore be considered together. The opposite is also imaginable and consequently fragments dealing with totally different issues should be separated. Additionally, not much informaion about the documents is available and so clustering is the right choice for this setting [6]. There are multiple clustering algorithms available. The most popular is *k-means*. Choosing the appropriate number of clusters k is not a trivial task since the results of k-means strongly depend on the initial selection of seeds [6]. In the evaluation, the number of clusters k is determined by using the elbow method (cf. [24]) whereupon for each k the average variability (within sum of squares) is computed ten times and the arithmetic mean over these ten runs is used in the elbow plot. In combination with the information from the previous analysis steps a reliable decision on the number of clusters is thus possible.

The evolved clusters define the new corpora, for example, if the number of clusters is 10, also 10 corpora are set up. In the **2nd analysis** word clouds, histograms or dendrograms (not displayed in the paper) can be computed per cluster. These last two steps (Model building & 2nd Analysis) can of course be iterated per cluster if the number of document fragments is very large.

At the end, the topic(s) of document fragments in one cluster can be derived based on which a characterization of the fragmented documents becomes possible. As a final step (**Wordlist generation**, **Sentence extraction**) wordlists are built (per cluster) and used to extract significant sentences from the fragmented documents in one cluster (cf. Algorithm 1; ↦RQ3). The sentence extraction enables the outline of potential requirements and implementation guidelines.

```
Data: clustered fragmented documents
Result: significant sentences per cluster
for each cluster do
    build a corpus consisting of all documents in the cluster;
    determine (unique) uni- and bigram keywordlists based on frequent terms;
    optional: let user edit these lists;
    perform POS tagging per sentence for documents in the corpus;
    extract sentences containing at least one word (unigram or bigram) present
      in the keywordlists
end
```

Algorithm 1: Determine significant sentences for each cluster

Algorithm 1 computes per cluster unigram and bigram wordlists. These consist of frequent terms using both *weightTf* and *weightTfIdf.* A user could refine these lists or add terms that might be of importance. Then sentences are tagged using the function *Maxent_Sent_Token_Annotator()* contained in the *R*-package *OpenNLP*[7]. If a sentence contains a word present in the wordlists it is saved for output and the user can view all extracted sentences per cluster in a .txt file.

Overall, the presented methodology combines existing text mining and analytical techniques with a novel document fragmentation approach. The following case studies will illustrate the applicability of the methodology to documents from different domains and show that it faciliates a sound understanding of the fragmented documents.

3 Description of Documents

3.1 Description of ISO/IEC documents

For the first case study, 13 documents from the ISO 27000 security standard family, i.e., a document composition treating a similar topic (in this case IT security), were selected. This selection consists of ISO 27000_2014, ISO 27001 – ISO 27005, ISO 27010, ISO 27011, ISO 27013, and ISO 27032 – ISO 27035. Document ISO 27000_2014 is an overview document that guides the reader through the more specific topics of the following documents, e.g., guidelines for cybersecurity in ISO 27032. It also contains a glossary with important general terms. Moreover,

[7] https://CRAN.R-project.org/package=openNLP

every document also encloses a collection of terms specifically important for this document. The documents contain between 42 and 136 pages.

A qualitative assessment of the documents was conducted based on an expert interview. The results are summarized in the following:

1. ISO 27001: Overview and vocabulary: $\frac{1}{3}$ is management-related, $\frac{2}{3}$ technical; document has a special role, i.e., it provides a general overview for the management; document contains description of overall process; the components are defined in the other documents
2. ISO 27002: Code of practice; describes actors and roles as well as different aspects of organization (cf. roles) / management (cf. information processing)
3. ISO 27003: Guidance of implementation
4. ISO 27004: Information security management system (ISMS) measurement
5. ISO 27005: Risk assessment, breakdown of risk assessment
6. ISO 27010: Information exchange inter-organizational, inter-sector / inter-organizational communication
7. ISO 27011: Instantiation for telecommunication
8. ISO 27013: Integrated implementation (document per se not so relevant)
9. ISO 27032: Cybersecurity
10. ISO 27033: Network Security
11. ISO 27034: Application Security
12. ISO 27035: Information Security, Incident Management

3.2 Description of Medical document

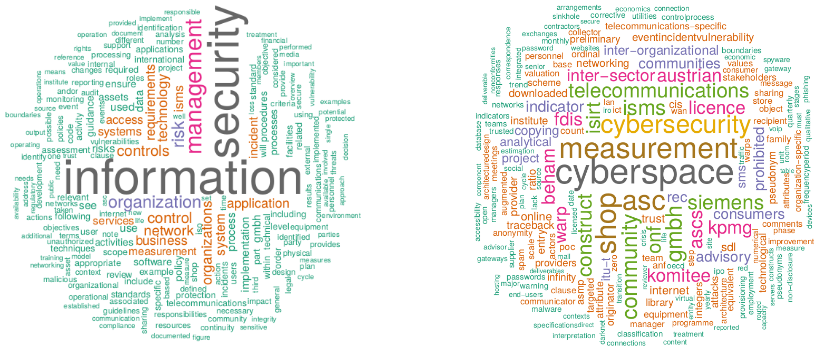
The document "Diagnosis and treatment of melanoma: European consensus-based interdisciplinary guideline" [8] consists of 14 pages and contains instructions on cutaneous melanoma diagnosis and treatment. Within a previous case study (cf. [3]) process models and constraints were manually resolved from the document. Therefore it will be possible to compare this manual outcome with the results of the application of the presented methodology. Selecting this document for a second case study is reasonable since it illustrates the variety of domains and document collections the methodology can be applied to (cf. Figure 2).

4 Evaluation

The methodology outlined in Sect. 2 is applied within three case studies. First, to security documents from the ISO 27000 family (cf. Sect. 3.1) and secondly on a medical document (cf. Sect. 3.2). The third case study as well as all results and figures can be downloaded[4].

4.1 Case Study 1: ISO Documents

As described in Sect. 3.1, 13 documents were chosen for the first case study. All documents are merged into one corpus *securityAll* and the previously described corpus transformations are applied. Following the methodology, depicted in Figure 1 in a **1st analysis** the most frequent terms are visualized using a word cloud with a maximum of 100 words and a histogram (taking the distribution over the documents into account), always applying *weightTf* as well as *weightTfIdf*. The results are displayed in Figure 3a and 3b.



(a) Word cloud for *securityAll*, *weightTf*

(b) Word cloud for *securityAll*, *weightTfIdf*

Fig. 3: Word clouds for *securityAll*

What can be recognized but is not surprising is that the terms `information` and `security` are frequent. Also the terms `management, iso/iec, isms, organization, risk` as well as `measurement, cyberspace, cybersecurity` and `telecommunications` occur quite often. Figure 4 shows the distribution of frequent terms (computed for *weightTf*) in each of the selected documents (one color per document). What can be observed is that `telecommunications` only shows up in ISO 27011 whereas `information` and `security` are present in each document. In Figure 5 *weightTfIdf* was used and the frequent terms are even more correlated to a specific document.

After this first analysis step, all documents are (semi-) automatically split into sections. For this **data preparation** step a R-Script was implemented which (per document)

1. extracts the table of contents via regular expressions; each section headline corresponds to one entry in a vector
2. searches the main part of the document for the section headlines
3. extracts the part in between the section headlines (inclusive headlines)
4. saves the fragments to separate txt files.

This script is tailored to the structure of the security documents, but its main idea (to perform fragmentation according to the table of contents) can be adopted to fit other document structures. Section headlines are saved twice since they are considered to contain important terms. The fragmentation level was set at section level for all documents in this case study, so in the end 202 fragments are obtained. For the subsequent steps, the introduction, as well as

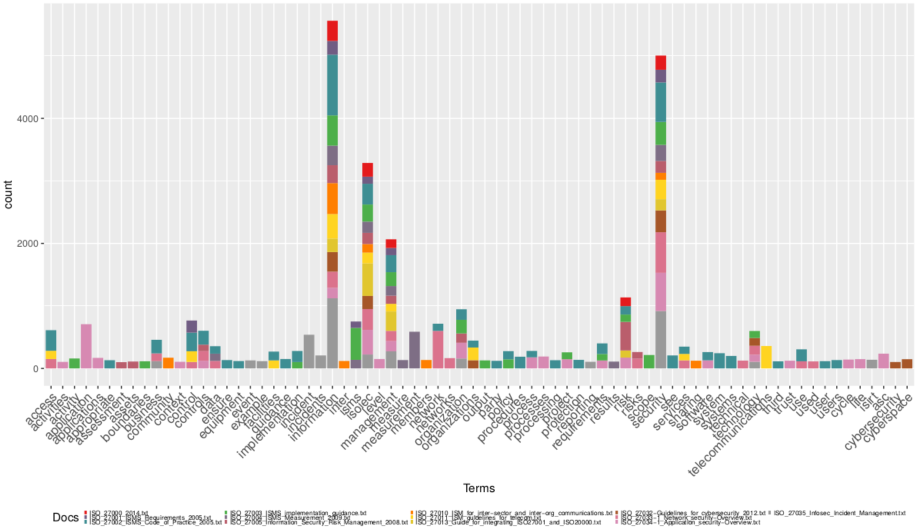


Fig. 4: Histogram for *securityAll*, *weightTf*

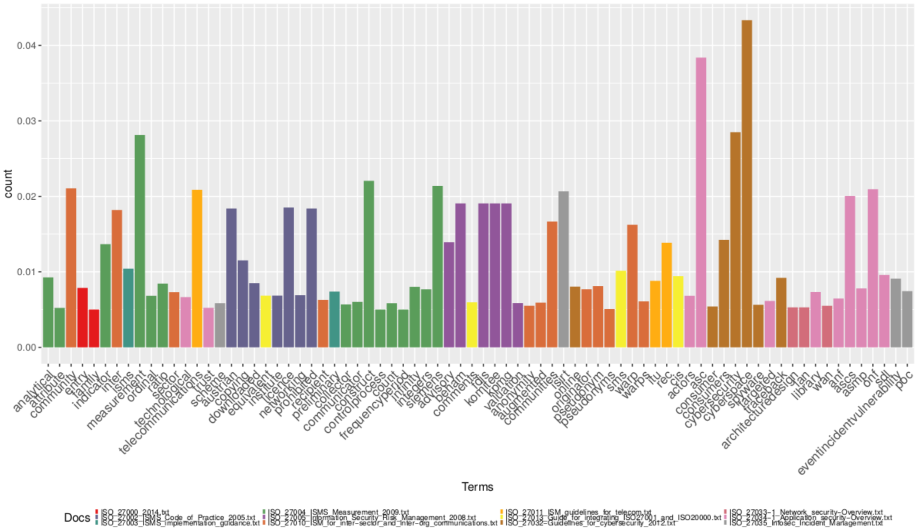


Fig. 5: Histogram for *securityAll*, *weightTfIdf*

the scope and terms and definitions sections of each document were not included since these fragments would increase the frequency of already frequent terms and would therefore cause noise which should be avoided. Thus the second corpus *securitySections* only contains 129 files.

After preprocessing the second corpus, word clouds and histograms are computed in order to see if the results have improved compared to the first analysis. Only when using *weightTfIdf* a change was noticed. In this case the terms `ISMS, inter-organizational, inter-sector, risk` and `community` are more recurrent than for corpus *securityAll*.

For **model building**, the number of clusters for k-means is determined. Therefore an elbow plot is used. For the document-term matrix *weightTfIdf* is applied and the distance measure is the cosine distance. The initial centers are selected randomly and as mentioned in Sect. 2 the average variability (within sum of squares) is computed ten times for each k but the plot (cf. Figure 6) includes only the arithmetic mean over these ten runs.

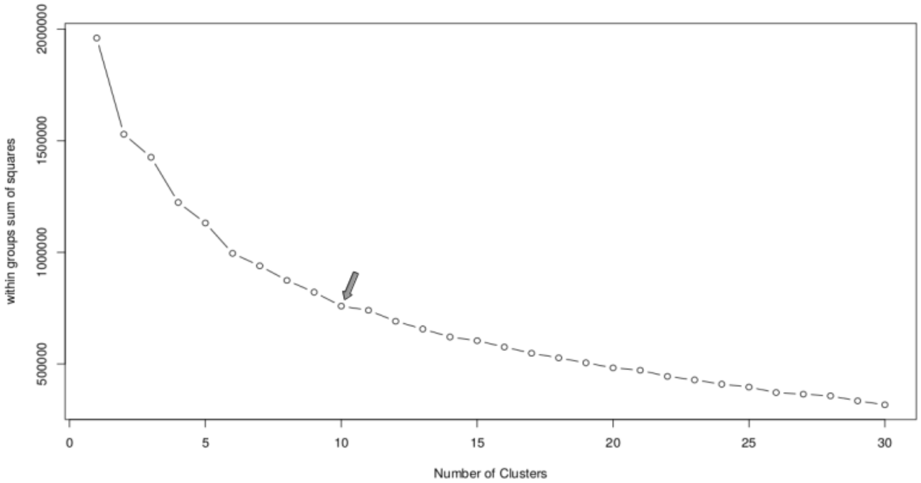


Fig. 6: Elbow plot for *securitySections*

According to Figure 6, 10 clusters are reasonable, so k-means is performed for $k = 10$. The resulting clusters contain between 7 and 25, in average 12.9 fragments. In a **2nd analysis** for each of these clusters word clouds and histograms are computed. Additionally, the **wordlist determination** is performed in order to **derive significant sentences**.

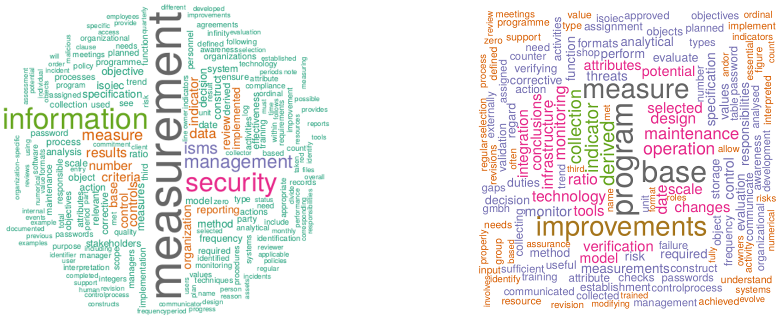
For showing the feasibility of the methodology the results for one randomly picked cluster are given in the following.

Example Cluster (Security) ($\mathcal{ECS}$): All 9 fragments of this cluster are put into a corpus *corpusECS* and after preprocessing *corpusECS* it can be recognized that the following terms stand out

- unigram word clouds (cf. Figure 7a and 7b): `measurement, information, security, isms, management, results, criteria, data, organi-`

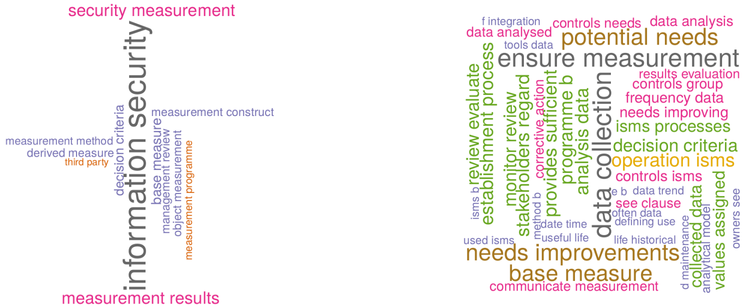
`zation, effectiveness, program, base, improvements, assurance, integration, verification, risk`

- bigram word clouds (cf. Figure 8a and 8b): `security measurement, measurement results, derived measure, base measure, ensure measurement, data analysis` and `data collection`.



(a) Word cloud for $\mathcal{ECS}$, weightTf (b) Word cloud for $\mathcal{ECS}$, weightTfIdf

Fig. 7: Unigram word clouds for $\mathcal{ECS}$



(a) Word cloud for $\mathcal{ECS}$, weightTf (b) Word cloud for $\mathcal{ECS}$, weightTfIdf

Fig. 8: Bigram word clouds for $\mathcal{ECS}$

Based on these results it can be concluded that the documents contained in this cluster treat the measurement and evaluation of ISMS, as well as topics on data collection, storage and handling of ISMS in general. Responsibility assignments are also covered in the fragments. This is checked by inspecting the fragments which are `5.information security measurement overview, 6.management responsibilities, 7.measures and measurement development, 8.measurement operation, 9.data analysis and measurement results reporting, 10.information security measurement programme evaluation and improvement, annex a, annex b` all contained in ISO 27004 and `annex e` from ISO 27003. The observed terms and deduced topics fit the description of ISO 27004 and 27003 in Sect. 3.1.

Now the overall topic of the fragments is known and two specific wordlists (one for unigrams and another for bigrams) can be acquired. This is done by resolving frequent unigrams respectively bigrams with function *freqTerms* implemented in the tm-package. Additionally, a domain expert could extend and refine these wordlists which can now be applied in order to figure out relevant phrases and sentences.

The wordlists for the selected cluster are

- Unigrams: `base, construct, control, criteria, data, decision, indicator, information, isms, management, measure, measurement, method, number, organization, reporting, reserved, results, review, rights, security, specification`
- Bigrams: `base measure, decision criteria, derived measure, frequency data, information security, management review, measure specification, measurement construct, measurement method, measurement results, object measurement, rights reserved, security measurement, siemens ag, third party`.

Three examples of the derived sentences determined by Algorithm 1 are

- "An organization should develop and implement measurement constructs in order to obtain repeatable, objective and useful results of measurement based on the information security measurement model."
- "The information security measurement programme and the developed measurement construct should ensure that an organization effectively achieves objective and repeatable measurement and provides measurement results for relevant stakeholders to identify needs for improving the implemented isms, including its scope, policies, objectives, controls, processes and procedures."
- "All relevant measures applied to an implemented isms, controls or groups of controls should be implemented based on the selected information needs.".

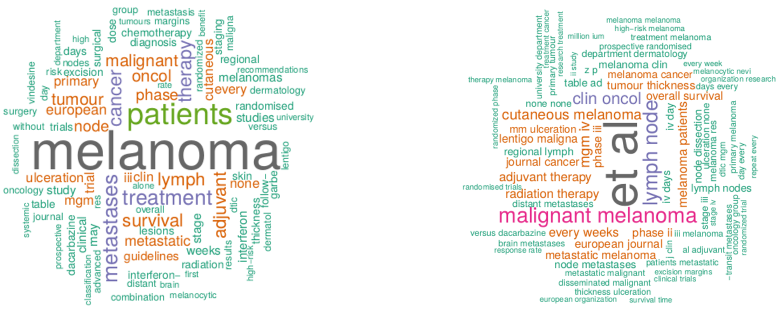
Implementation instructions are clearly contained in these sentences. So it is possible to figure out requirements (semi-) automatically by combining standard text mining tools with fragmentation and clustering of documents.

The ISO documents have significant section titles and so the advantage of the approach concentrates more on the automatic grouping of fragments, the (semi-) automatic deduction of wordlists and the extraction of relevant sentences than giving a first insight of the content.

4.2 Case Study 2: Medical document

As described in Sect. 3.2, the medical document differs from the ISO document collection. There is only one short document (14 pages) available compared to the ISO documents (≥ 42 pages). Moreover, it has a different structure, e.g., no table of contents is included. So, fragmentation based on the table of contents is not possible here. Instead, one has to figure out appropriate regular expressions for being able to perform an automatic fragmentation.

After importing the document into the corpus *corpusMedical*, transformations like converting all words to lower case and removing customized words are applied. Performing the **1st analysis** step of the methodology generated the word clouds depicted in Figure 9. Applying *weightTfIdf* is not reasonable here, since *corpusMedical* contains only one document. Examples of frequent unigrams and bigrams are `melanoma, patients, metastases, treatment, cancer, therapy, et al, lymph node, malignant melanoma, cutaneous melanoma`. Having a closer look at the remaining terms gives a good impression of the content of the document. One could verify this by e.g., reading the abstract.



(a) Unigram word cloud

(b) Bigram word cloud

Fig. 9: Word clouds for the medical document

For the **data preparation** the document is fragmented. As mentioned before, the document does not contain a table of contents, so the fragmentation is issued via suitable regular expressions. Splitting the document into subsections results in 35 partial documents while fragmentation on section level produces just 7 partial documents. For this analysis the subsection level is chosen in order to produce a fine-granule fragmentation.

As in the first case study an elbow plot for determining the number of clusters is used (cf. Figure 10; **model building**). Based on this plot $k = 6$ is chosen for $k-$means clustering and the clusters contain between 3 and 9 fragments. Like before, the results of one randomly picked cluster are given below.

Example Cluster (Medical) ($\mathcal{ECM}$): Following the methodology outlined in Sect. 2 a **2nd analysis** is issued on the five documents contained in the cluster. For this purpose the documents are merged into a corpus *corpusECM*. Subsequently, corpus transformations are again applied on this corpus and word clouds are computed (cf. Figure 11 and 12). Here, *weightTfIdf* can be used again since the corpus contains more than one document.

In contrast to the primary word clouds a more detailed description of the documents in $\mathcal{ECM}$ seems to be possible. Here, the frequent terms are not only `metastases` and `therapy` but also different types of metastases like `skin metastases, distant metastases, bone metastases, brain metastases`. The documents in the cluster also seem to contain therapy and treatment sug-

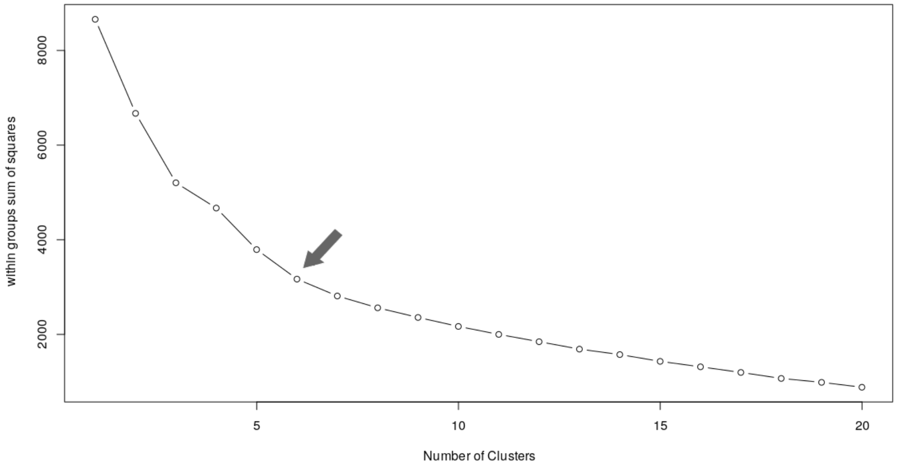
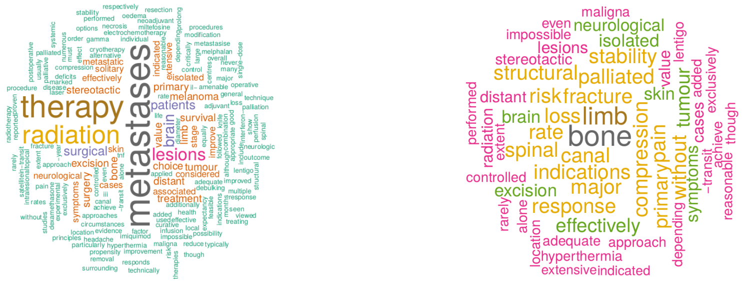


Fig. 10: Elbow plot for the medical document

gestions (cf. `patients, survival, radiation, surgical, rate, indications, pain, stability, treatment choice, effectively palliated`).



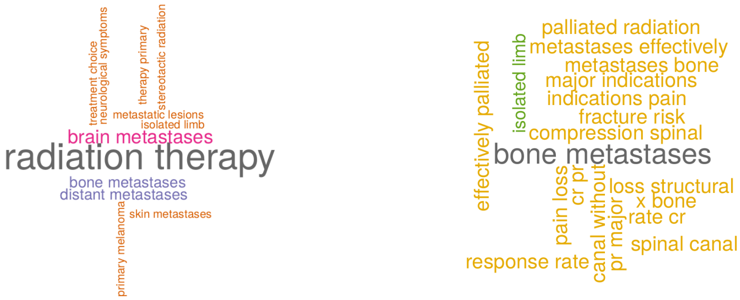
(a) Unigram word cloud, weightTf

(b) Unigram word cloud, weightTfIdf

Fig. 11: Unigram word clouds for $\mathcal{ECM}$

Keyword lists are generated and used to *extract sentences* resulting in

- "For multiple lesions on a limb, isolated limb perfusion with melphalan ± tumour necrosis factor (TNF) has palliative value."
- "In stage iii patients with satellite/in-transit metastases the procedure can be curative, as indicated by the reported 5- and 10-year survival rates of 40% and 30%, respectively."
- "Even though excision is the treatment of choice for lentigo maligna, radiation therapy may achieve adequate tumour in-transit metastases, which



(a) Bigram word cloud, weightTf

(b) Bigram word cloud, weightTfIdf

Fig. 12: Bigram word clouds for $\mathcal{ECM}$

are too extensive for a surgical approach, may be effectively controlled by radiation therapy alone."

Without an expert interview a qualitative assessment of the derived terms and sentences is not possible right away, but reading through the partial documents in the cluster (`3.9 skin metastases, 3.10 distant metastases, 4.1 primary melanoma, 4.4 bone metastases, 4.5 brain metastases`) can provide an evaluation for the feasibility of the clustering, i.e., if the topic of the documents in the cluster is derived correctly. This is the case for $\mathcal{ECM}$.

Nevertheless, in order to be able to compare the significance of the derived sentences per cluster, we searched for the cluster with the partial document containing a manually derived subprocess (cf. Figure 13). In fact, the methodology determined the following sentences for this cluster:

"There is considerable variation in follow-up approaches and few data to support them. In stages I-II melanoma, the intent is to detect loco-regional recurrence early so that the frequency of follow-up examination is usually every 3 months for the first 5 years, whereas for the 6-10th year period investigations every 6 months seem to be adequate."

The second sentence represents a constraint similar to the one depicted by the subprocess in Figure 13 which leads to the conclusion that it is possible to extract constraints (respectively requirements) using the presented approach. Due to lack of space the third case study based on a selection of privacy documents is not in the paper but can be downloaded[4]. Based on the methodology it was possible to group the fragmented documents into clusters having similar topics. For example one cluster treats the legal aspects of privacy and their impact on the society and companies. For this cluster the terms `rights, fundamental, article, impact, human, protection, sources, executive, fundamental rights, impact assessment, european union` were derived using word clouds and histograms. Based on these frequent terms, sentences were found that contain implementation instructions.

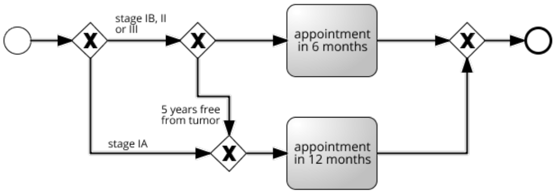


Fig. 13: BPMN Model: Sonography Aftercare Subprocess (cf. Fig. 4 in [3])

5 Limitations and Application Scenarios

The evaluation section has demonstrated the applicability of the methodology on several security and one medical document. Many more use cases encountering other domains and document sets are imaginable. But how can organizations and institutions benefit from these results, what limitations might emerge from the methodology and what are application scenarios?

By now, the main focus for visualizing frequent terms is on word clouds and histograms, but for a more in-depth understanding of the documents further techniques might be useful. Dendrograms, i.e., hierarchical clustering of terms are only one possible technique. Another one is to perform association analysis for selected frequent terms and to display the results by, e.g., network maps.

Even though the methodology delivered a reasonable grouping of fragmented documents as well as significant sentences it is necessary to evaluate the completeness of these sentence collections per cluster. It is likely that not every important term is represented by a frequent unigram or bigram. For evaluating the completeness and correctness domain experts should additionally be consulted.

Therefore, in combination with domain knowledge the methodology can extract main characteristics of a series of documents related to planning and implementing of regulatory documents and best practices in existing standards and guidelines. This makes it easier for organizations to accelerate the installation and maintenance of compliance guidelines and related processes, shorten the consulting procedure, and reduce the overall implementation costs. Nevertheless, the methodology does not provide an interpretation of how guidelines should be brought into action in a specific setting.

Another limitation of the approach arises from “hidden” information in pictures and tables. One solution is to use tools that are able to parse such information to make it available for an analysis in *R*.

Most requirements elicitation approaches demand additional knowledge (cf. Sect. 6). Here, additional knowledge consists of, e.g., overview documents or glossaries, and might decrease implementation effort and increase the output quality during the clustering stage where document fragments are grouped together. Nevertheless, such additional knowledge is not mandatory for this methodology.

6 Related Work

This work touches different areas, i.e., requirements engineering, deriving process-related information from text, and text mining. Thus related work from these areas will be discussed in the following.

Requirements engineering is a broadly investigated field where requirements elicitation constitutes one of the phases in the engineering process. Out of the multiple frameworks for requirements engineering, some approaches suggest (semi-)automated requirements elicitation/identification techniques. The survey presented in [18] has investigated in how far the requirements identification phase is supported in an automated manner. There are some approaches that support or envision (semi-)automatic requirements identification. The majority of these approaches requires additional knowledge, e.g., an ontology. While this is promising for the presented approach as well, the work presented in this paper does not assume additional knowledge. The only automatic approach that does not require additional knowledge (acc. to [18]) is [1]. In contrast to the methodology presented in the paper at hand, [1] assume short forum posts as input. [21] conducted a systematic review on security and privacy requirements elicitation approaches and concluded that there is only little support for automated requirements elicitation. Overall, the approach presented in this work can be seen as support for requirements elicitation and can be employed with any of the requirements engineering frameworks. After extracting requirements, contextualization, i.e., the interpretation of requirements is often useful. An approach using predefined templates is presented in [14].

Implementing requirements contained in regulatory documents often correlates with determining process models and constraints that restrict the implementation. Process model discovery from textual sources is envisaged by several existing approaches (cf. e.g., [2, 7, 9, 10, 19, 20]), but how to derive these from natural language documents is still an open question (cf. [15]). The reason is that the mentioned approaches have limitations and partly strict requirements on the textual information. The approach presented by [7], for example, requires "the description to be sequential and to contain no questions and little process-irrelevant information.". This, in turn, restricts the outcome, i.e., the process models, as well, as real-world processes are often not purely sequential, but involve further patterns such as decisions. For requirements elicitation, in addition, information such as on frequently executed activities can be useful as well (cf. medical case study in Section 4.2). Opposed to existing approaches, this work aims at neither imposing restrictions on the text of interest nor on the outcome.

[5] outline the prototypical text mining process in R including common corpus transformations as mentioned in Sect 2. Employed weighting schemes for term-document matrix construction follow best practices as discussed in [17] and [23]. Chunking of long documents in shorter fragments has been proposed by [4] in the context of stylometry of texts. However, their approach focused on better visualization (by having more data points) but ignored the semantic fragmentation we highlight in this paper (where connected fragments and paragraphs capture coherent semantic concepts). Clustering text documents in

R was discussed in [12] but focused on clustering techniques and distance measures. We instead use clustering as an explanatory guiding technique that needs specific treatment; e.g., for finding a good number of clusters we used an elbow plot [24] which is useful as an alternative to silhouette plots [22]. The combination of quantitative (classical text mining techniques) and qualitative (expert interview) methods provides hereby a unique contribution for the specific domains considered in this paper.

7 Conclusion and Future Work

This paper outlined a methodology for (semi-) automatically characterizing compliance and regulatory documents by applying well-known text mining and clustering methods like resolving frequent terms or *k*-means. The evaluation has demonstrated how and to what extent a user can be supported in implementing requirements based on these types of documents.

Future work will encounter conducting user studies in order to further evaluate the usefulness of the methdology for domain experts. The inclusion of topic models in order to improve the results of the presented methodology will be another aspect as well as to try POS tagging for resolving process elements like actions or roles since this assists in implementing requirements more precisely.

References

1. Castro-Herrera, C., Duan, C., Cleland-Huang, J., Mobasher, B.: A recommender system for requirements elicitation in large-scale software projects. In: Symposium on Applied Computing. pp. 1419–1426 (2009)
2. Deeptimahanti, D.K., Babar, M.A.: An automated tool for generating uml models from natural language requirements. In: Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering. pp. 680–682. ASE '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/ASE.2009.48
3. Dunkl, R., Fröschl, K.A., Grossmann, W., Rinderle-Ma, S.: Assessing medical treatment compliance based on formal process modeling. In: USAB 2011 – Information Quality in eHealth. pp. 533–546. Springer (2011)
4. Feinerer, I.: An introduction to text mining in R. R News 8(2), 19–22 (2008), http://CRAN.R-project.org/doc/Rnews/
5. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. Journal of Statistical Software 25(5), 1–54 (2008)
6. Feldman, R., Sanger, J.: The text mining handbook : advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge; New York (2007)
7. Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: Int'l Conf. Advanced Information Systems Engineering. pp. 482–496 (2011)
8. Garbe, C., Peris, K., Hauschild, A., Saiag, P., Middleton, M., Spatz, A., Grob, J.J., Malvehy, J., Newton-Bishop, J., Stratigos, A., et al.: Diagnosis and treatment of melanoma: European consensus-based interdisciplinary guideline. European journal of cancer 46(2), 270–283 (2010)

9. Ghose, A., Koliadis, G., Chueng, A.: Rapid Business Process Discovery (R-BPD), pp. 391–406. Springer Berlin Heidelberg, Berlin, Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-75563-0_27
10. Gomez, F., Segami, C., Delaune, C.: A system for the semiautomatic generation of e-r models from natural language specifications. Data & Knowledge Engineering 29(1), 57 – 81 (1999), http://www.sciencedirect.com/science/article/pii/S0169023X98000329
11. Hill, T., Lewicki, P.: Statistics : methods and applications : a comprehensive reference for science, industry, and data mining. Tulsa, Okla. : StatSoft ; [United Kingdom] : [StatSoft Ltd.] (2006)
12. Hornik, K., Feinerer, I., Kober, M., Buchta, C.: Spherical k-means clustering. Journal of Statistical Software 50(10), 1–22 (2012), http://www.jstatsoft.org/v50/i10
13. Bank for International Settlements: Basel 3: International framework for liquidity risk measurement, standards and monitoring (2010)
14. Koliadis, G., Desai, N.V., Narendra, N.C., Ghose, A.K.: Analyst-mediated contextualization of regulatory policies. In: Services Computing (SCC), 2010 IEEE International Conference on. pp. 281–288. IEEE (2010)
15. Leopold, H.: Natural language in business process models. Springer (2013)
16. Ltd, I.G.: it governance: Iso 27001 global report (2016), http://pribatua.org/wp-content/uploads/2016/08/ISO27001-Global-Report-2016.pdf
17. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
18. Meth, H., Brhel, M., Maedche, A.: The state of the art in automated requirements elicitation. Information & Software Technology 55(10), 1695–1709 (2013), https://doi.org/10.1016/j.infsof.2013.03.008
19. More, P., Phalnikar, R.: Generating uml diagrams from natural language specifications. International Journal of Applied Information Systems 1(8), 19–23 (2012)
20. Omar, N., Hassan, R., Arshad, H., Sahran, S.: Automation of database design through semantic analysis. In: Proceedings of the 7th WSEAS international conference on Computational intelligence, man-machine systems and cybernetics, CIMMACS. vol. 8, pp. 71–76 (2008)
21. Rinderle-Ma, S., Ma, Z., Madlmayr, B.: Using content analysis for privacy requirement extraction and policy formalization. In: Enterprise Modelling and Information Systems Architectures. pp. 93–107 (2015)
22. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)
23. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
24. Thorndike, R.L.: Who belongs in the family? Psychometrika 18(4), 267–276 (1953)