Towards Identifying GDPR-Critical Tasks in Textual Business Process Descriptions

Nake, Leonard; Kuehnel, Stephan; Bauer, Laura; Sackmann, Stefan (2023): **Towards Identifying GDPR-Critical Tasks in Textual Business Process Descriptions**, In: M. Klein, D. Krupka, C. Winter., V. Wohlgemuth (Eds.): INFORMATIK 2023, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023, pp. 1895-1908.

Acknowledgements

The project on which this study is based was funded by the German Federal Ministry of Education and Research under grant number 16KIS1331. The responsibility for the content of this publication lies with the authors.

Martin Luther University Halle-Wittenberg
Universitaetsring 3
06108 Halle (Saale)
http://informationsmanagement.wiwi.uni-halle.de

M. Klein, D. Krupka, C. Winter., V. Wohlgemuth (Hrsg.): INFORMATIK 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023 1895

Towards Identifying GDPR-Critical Tasks in Textual Business Process Descriptions

Leonard Nake [1], Stephan Kuehnel [1], Laura Bauer [1], and Stefan Sackmann [1]

Abstract: Complying with data protection regulations is an essential duty for organizations since violating them would lead to monetary penalties from authorities. In Europe, the General Data Protection Regulation (GDPR) defines personal data and requirements for dealing with this type of data. Hence, organizations must identify business activities that deal with personal data to establish measures to fulfill these requirements. Especially for large organizations, a manual identification can be labor-intensive and error-prone. However, textual business process descriptions, such as work instructions, provide valuable insights into the data used in organizations. Therefore, we propose a first approach to automatically identify GDPR-critical tasks in textual business process descriptions. More specifically, we use a supervised machine learning algorithm to automatically identify whether a task deals with personal data or not. A first evaluation of our approach with a dataset of 37 process descriptions containing 509 activities demonstrates that our approach generates satisfactory results.

Keywords: Legal Compliance, General Data Protection Regulation (GDPR), Business Process, Task Identification

1 Introduction

Protecting the IT infrastructure is an essential task for companies since information security incidents become more frequent and the costs of managing and mitigating breaches have been growing over the years [Ac21]. Due to this phenomenon, legislators and companies define overarching requirements for IT security that companies must comply with. For instance, Article 32 (1) of the EU General Data Protection Regulation (GDPR) requires an organization to implement appropriate technical and organizational measures to ensure compliance with the protection goals of confidentiality, integrity, availability, and resilience when processing personal data. To fulfill this requirement, technical precautions (e.g., encryption and pseudonymization of personal data) and procedural configurations (e.g., activities and controls to ensure compliance in business processes) become necessary [Kü21]. Therefore, compliance with IT security requirements is a cost-intensive task [La15], which makes it essential to determine where potential IT security measures have to be applied since economic efficiency in this context

[1] Martin Luther University Halle-Wittenberg, Chair for Information Management, Universitaetsring 3, 06108 Halle (Saale), {leonard.nake, stephan.kuehnel, laura.bauer, stefan.sackmann}@wiwi.uni-halle.de,
https://orcid.org/0000-0001-8324-5641, https://orcid.org/0000-0002-6959-9555,
https://orcid.org/0000-0001-8911-0879, https://orcid.org/0000-0002-3370-6785

doi: 10.18420/inf2023_191

is important [CCR04].

One of the requirements established by the GDPR is defined in Article 32 (2): "In assessing the appropriate level of security account shall be taken in particular of the risks that are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to personal data transmitted, stored or otherwise processed." This requirement makes it necessary for organizations to determine where personal data is dealt with in their business processes. These business processes consist of sets of activities performed in the organization [We19]. Only when the activities dealing with personal data are identified, adequate measures can be established. For small organizations with only a few business processes, analyzing each activity manually might be a viable solution. However, larger organizations agglomerate vast numbers of business processes since even single projects can result in the creation of hundreds or more new business processes [BRU00]. Each of these business processes can contain hundreds of activities of different departments and legal entities [RMv09]. Hence, a manual analysis of all of these business processes regarding the use of personal data becomes a difficult and labor-intensive problem.

One widely used possibility to store information about business processes is in textual business process descriptions. They contain valuable information that could be utilized in such an analysis since organizations usually maintain hundreds of textual process descriptions [LvR18]. For this reason, some approaches automatically analyze textual process descriptions to gain insights into the business processes of organizations [LvR18], [FMP11]. However, to the best of our knowledge, there is no approach that analyzes such textual business process descriptions to gain insights into the use of personal data.

To address this research gap, we propose an approach that analyzes textual business process descriptions to gain insights into the transmission, storage, or processing of personal data in business processes. More precisely, the proposed approach analyzes the natural language inside the textual business process descriptions to automatically classify each task as either GDPR-critical or GDPR-uncritical. A task is seen as GDPR-critical if it deals with personal data defined by Article 4 (1) of the GDPR. Hence, we raise the following research question:

How to automatically identify GDPR-critical tasks in textual business process descriptions?

To address this research question, we base our research design on the method proposed by Leopold et al. [LvR18]. Our research contributions are threefold: we introduce three features, train a model, and conduct an evaluation. Our study shows that the proposed approach with the current dataset achieves a F1-score of 0.81. This paper is structured in the following way: In section 2, we discuss the theoretical background and related work. In section 3, we present our research design. Section 4 deals with the conceptual approach as it describes the proposed concept in detail. Section 5 contains an evaluation of the conceptual approach. Section 6 concludes the paper.

2 Theoretical Background and Related Research

Our paper is thematically located in the area of business process compliance. Compliance refers to the adherence to rules, i.e., acting in accordance with applicable regulations, which can originate from various sources, such as laws, directives, standards, etc. [Ra12]. More specifically, business process compliance deals with identifying, formalizing, implementing, checking, analyzing, and optimizing compliance requirements before, during, or after the execution of business processes [SKS18]. An important source of regulations on the security of information and data in Europe, which we focus on in this study, is the GDPR. Data protection laws in Europe have long been inconsistent and the development of the GDPR has addressed this issue [Se20]. The regulation provides individuals with more protection and control over their personal data in light of new technological developments [To20], [Se20]. To this end, Article 4 (1) EU GDPR writes the following about personal data: "'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;" Since its entry into force, violations of, e.g., Article 32 EU GDPR ("security of processing") in conjunction with Article 83 ("general conditions for imposing administrative fines") have been punishable by fines of up to EUR 20 million or 4% of the total worldwide annual turnover of the previous fiscal year. Thus, the level of potential sanctions increased dramatically compared to previously applicable rules, and so did companies' needs for security to protect them against compliance violations resulting from the GDPR. Since it is not an easy task to keep track of all information and data processing activities of a company, our approach starts exactly at this point and makes use of task analysis with Natural Language Processing (NLP) technologies and supervised machine learning (SML). While NLP uses computer-based methods to understand, produce and learn human language content [HM15], SML uses labeled datasets to train algorithms for classifying data or accurately predicting outcomes [Li11].

This paper relates to two major streams of research. Firstly, technologies for identifying requirements and checking completeness in the context of privacy policies. Secondly, the application of NLP technologies in the context of business process analysis. The first stream of research can be divided into two types of approaches: identifying requirements of privacy policies and completeness checking of privacy policies. Regarding the first type, Caramujo et al. [Ca19] propose a domain-specific language for specifying privacy policies in the context of mobile and web applications. This language allows policy authors to define a privacy policy as a set of declarative statements. They apply this language to support the analysis and comparison of policies from different companies. Pullonen et al. [Pu19] present a multi-level model as an extension of BPMN (Business Process Model and Notation). This enables the user to visualize, analyze, and communicate the characteristics of privacy policies in business processes. Kumar and

Shyamasundar [KS14] use information flow controls as a means to specify and enforce privacy policy requirements. Although some of these papers also use business processes to identify privacy policy requirements, the overall goal is different since we try to identify tasks that deal with personal data. Additionally, they are not strictly based on the GDPR. Regarding completeness checking, Torre et al. [To20] and Rahat et al. [RLT22] propose approaches that check the completeness of privacy policies against the GDPR. In summary, the approaches from the first research stream deal with similar problems. However, there is no approach that addresses our research question. The second stream of research is task analysis utilizing natural language texts in business process management. In this field, process-related textual documents are analyzed to gain insights about the tasks of organizations [RRM21]. Relevant work in this area extracts process models from textual business process descriptions [FMP11], compares textual descriptions with process models [vLR17], identifies candidates for task automation [LvR18], or extracts relevant task content aspects from textual task descriptions [RRM21]. Our research can be seen as a bridge between the two main research streams. It uses methods from the task analysis in business process management domain to solve a problem regarding privacy policies, more specifically, the GDPR. To the best of our knowledge, there is no existing approach for automatically identifying GDPR-critical tasks using textual business process descriptions.

3 Research Design

This research is part of a larger Design Science Research (DSR) project [Kü21] following the procedure proposed by Vaishnavi and Kuechler [VK15] with the goal to create a method that automatically identifies GDPR-critical tasks in textual business process descriptions. In this paper, we present the results of the first iteration, where we took an already existing method, applied it to our identified problem, and evaluated the results. In this study, we base our research design on an existing approach by Leopold et al. [LvR18]. In their research, tasks are identified by analyzing the natural language in textual business process descriptions following the approach by Friedrich et al. [FMP11]. Subsequently, the analyzed textual business process descriptions are used to compute a set of features. Using these features, the identified tasks are then classified into manual, user, or automated tasks using SML. This is done to identify candidate tasks for robotic process automation. We follow the overall structure of this three-step approach by Leopold et al. [LvR18] but adapt the second and the third step for our approach to answer our research question. Figure 1 shows a visualization of the proposed approach. Firstly, a textual description of a business process is parsed to identify the linguistic entities in the text, such as the verbs and objects that describe a task. This step results in an annotated textual process description with identified tasks and their respective linguistic entities. Secondly, the linguistic entities of the annotated textual process description are analyzed to compute features for our model. More specifically, the object and data item of the identified task are extracted as features and the customer relation of the task is calculated. This step produces a table of the tasks as well as their features. Thirdly, the classification is performed using the features. We use a support vector machine (SVM), a SML approach,

to classify tasks as GDPR-critical or GDPR-uncritical based on manually classified training data. The output of this step is a list of classified tasks.

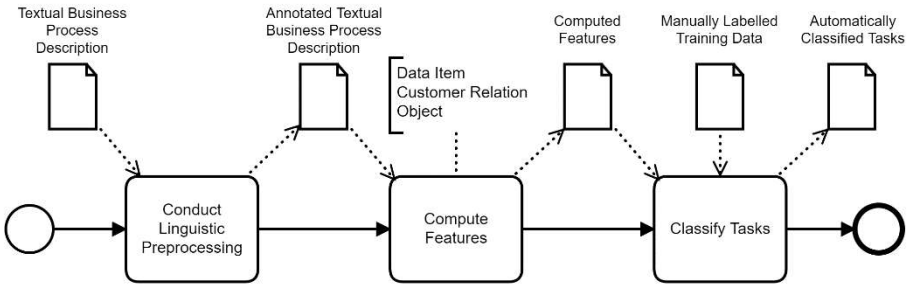

Fig. 1: Research Design

4 Concept of the Approach

4.1 Idea for the Approach

The idea behind our approach is to automatically identify an organization's tasks that deal with personal information as defined by the GDPR. This is an essential step in organizations' effort to comply with data protection regulations since possible sources of sensitive information must be identified to be able to apply adequate security measures. There are tasks where this is a simple problem that could be solved by only analyzing the task name, e.g. "Receive Customer Data" (Technical University of Berlin). However, many tasks in business processes have generic names, e.g. "Process Data" (Private Companies). To be able to accurately classify these tasks, it is necessary to include more context of the business process. Ideas for including additional information could be to analyze the data object of a task. For instance, if the data object is a "medical file" (Eindhoven University of Technology) then the task should be classified as GDPR-critical. If the data object is a "job description" (BPMN Handbook) on the other hand, the task should be classified as GDPR-uncritical. Another idea is to analyze the resource that executes the task. If the task "Process Data" is executed by the "Engineering Department" (Humboldt University of Berlin), it is unlikely that it deals with personal data. However, if the same task is executed by the "Customer Service" (Technical University of Berlin), the data referred to in the task name likely contains personal information.

4.2 Linguistic Preprocessing of Textual Business Process Descriptions

The linguistic preprocessing step is based on the work of Friedrich et al. [FMP11] excluding the actual creation of business process models. This step takes a textual business process description as an input and extracts verbs, objects, roles, and data items for each identified task. To achieve this, it uses NLP techniques such as the Stanford Parser [KM03] and VerbNet [Sc05]. The first step in the linguistic preprocessing is the application of the Stanford Parser that detects the part of speech of each word and the grammatical relations between words, also called part-of-speech tagging. It takes a sentence in natural language as input and identifies verbs as well as subjects and objects to which these verbs relate. The approach of Friedrich et al. [FMP11] uses the output of the tagging to automatically extract tasks consisting of a verb, an object, the executing role, and data items used.

4.3 Computation of Features

Given the ideas formulated in a prior section, we propose the following features for our model.

- **Data item (categorical).** The *data item* feature is categorical and relates to the name of the data item used in a task in the business process. We use data item as a synonym of the term data object defined in BPMN. Choosing the data items as a feature in our classification is a logical choice since we are interested in the data processed in each task. The rationale behind this feature is that there might be data items that almost always contain GDPR-critical data. Examples of such data items are “customer data” or “patient file”. But there also might be data items like “job description” that are likely to contain uncritical data since such documents tend to be publicized by the organization.
- **Customer relation (related/unrelated).** The *customer relation* feature is a binary feature that reveals whether the respective task has any relation to a customer. This is important because it allows the classification of tasks with ambiguous names. For instance, it is hardly possible to correctly classify a task called “enter data”, when only considering the name. However, if it can be determined that the data is from customers, it is possible to classify the task as GDPR-critical. Such a determination could be made by examining the context of each task. For instance, when the resource executing the task is the customer service, the processed data is likely from customers.
- **Object (categorical).** The *object* feature is categorical and relates to the grammatical object used in the name of a task. The idea behind this feature is that there are objects likely associated with the type of data used in a task. For instance, tasks with the objects “order” or “invoice” have an increased probability to deal with personal data. On the other hand, it is unlikely that tasks with certain objects, such as physical products (e.g., bicycles), deal with personal data.

The *data item* and *object* features are obtained during the linguistic preprocessing step while the *customer relation* feature is obtained through further analysis. The computation is done by analyzing the natural language used in the identified tasks and the relevant sentence in the process description. Should they contain the word customer or a synonym (e.g. patient), the *customer relation* feature has the value: *related*.

4.4 Classification of Tasks

Finally, we use a model to classify tasks from unseen business process descriptions as either GDPR-critical or GDPR-uncritical. Such a classification is not a trivial task as the context of each task varies. For instance, the task "send data" can hardly be correctly classified using a single feature. It is therefore necessary to combine the features from the previous chapter and apply a SML algorithm. A SML algorithm analyzes the training data and infers a function that is subsequently used to map new observations [Ru10]. As we are dealing with categorical variables where no ordinal relationship exists, we apply one-hot encoding. We then use a linear SVM [CV95] since it can deal with a small dataset, has a low risk of overfitting, and scales well. Because of this, SVMs are often used in text categorization problems to classify documents into predefined categories because they perform well in this particular set of problems [GDS19]. Other authors used SVMs for very similar text classification problems [Jo05], [LvR18], [TK01]. As input for our SVM, we provide manually labeled tasks as well as their computed features to train the model.

5 Dataset

We use the dataset of textual process descriptions introduced by Friedrich et al. [FMP11] with two differences. Firstly, we do not use all 47 process descriptions from the original dataset since 14 process descriptions from one source have a low language quality. Secondly, we expanded the dataset by one source of four textual business process descriptions from two companies. All of the 509 tasks were classified as either GDPR-critical or GDPR-uncritical by two researchers. If the classifications differed, a third researcher resolved conflicting labels. Some of these classifications were only possible by taking the whole business process description into account. For example, a task called "store information" was classified as GDPR-critical because the term "information" referred to the health information of a hospital patient. While most tasks were classified with confidence, there were some tasks where the classification of the first and second researcher differed. For instance, in a business process about receiving customer data, there was disagreement about a task called "initiate measures" that is executed in the case of unusual occurrences. One researcher argued that the mere initiation of measures does not involve personal data. The other researcher thought that since these measures are linked to customer data and cannot be initiated without directly referring to this customer data, the task should be classified as GDPR-critical. The third researcher agreed with the second researcher and the task was classified as GDPR-critical.

Table 1 shows the characteristics of the dataset. It should be noted that the identified tasks in Table 1 are not the direct result of the software tool by Friedrich et al. [FMP11]. Some of the identified tasks only consisted of a single verb or did not contain work to be performed and therefore violated the definition of an activity. Such tasks were excluded from the dataset. However, due to this exclusion, there is some subjectivity regarding the number of identified tasks. The textual business process descriptions of each source differ in several ways. Especially the length of the descriptions and sentences distinguishes the sources from each other. Another important difference is that the amount of GDPR-critical tasks varies greatly in the sources. This is due to the variety of business processes in the dataset. For instance, creating a job description (BPMN handbook) does not deal with personal data. The opposite is the reception of customer data (private companies) where almost all tasks explicitly deal with personal information about customers.

Type	Sources	D	S	SL	CT	UT
Academic	Humboldt University of Berlin	4	10.0	18.1	6	51
	Technical University of Berlin	2	34.0	21.2	37	37
	Queensland University of Technology	8	6.1	18.3	38	30
	Eindhoven University of Technology	1	40.0	18.5	23	21
Textbook	BPMN Handbook	3	4.7	17.0	3	14
	BPMN Guide	6	7.0	20.8	14	43
Industry	Vendor Tutorials	4	9.0	18.2	19	24
	inubit AG	4	11.5	18.4	15	40
	BPM Practitioners	1	7.0	9.7	6	1
	Private Companies	4	26.8	25.7	63	24
Total		37	15.6	18.6	224	285

Tab. 1: Details about the Dataset.

Legend: D = Number of textual business process descriptions, S = Average number of sentences, SL = Average sentence length in words, CT = Number of GDPR-critical tasks, UT = Number of GDPR-uncritical tasks, BPMN = Business Process Model and Notation, BPM = Business Process Management.

6 Evaluation

6.1 Description of the Evaluation

In a former section, we described the proposed approach on a conceptual level. In this section, we will evaluate this concept using our dataset by testing whether it can reliably predict the GDPR-criticality of tasks in unseen textual business process descriptions. The approach was implemented in Java using the prototype of Friedrich et al. [FMP11] and

the machine learning library Weka [Ha09]. We conducted a 5-fold cross-validation [HTF09] with a 60/20/20 split in training, validation, and test data with our dataset. Although we present our results by showing correctly and incorrectly classified tasks, the cross-validation is done on the business process descriptions to ensure that tasks from the same business process cannot be in the training data and the test data at the same time since this would lead to information leakage. To evaluate our approach as well as the proposed features, we used three configurations. Firstly, we trained only on the *data item* feature. Secondly, we trained on the *data item* and *customer relation* features. Thirdly, we trained on the *data item, customer relation* and *object* features. The quality of the configurations is measured using of precision, recall, and F1-score.

6.2 Results

Class	Metric	Data Item (DI)	DI & Customer Relation (CR)	DI & CR & Object
Tasks	Correct	63	86	86
	Incorrect	43	20	20
GDPR-critical	Precision	0.90	0.83	0.84
	Recall	0.29	0.83	0.81
	F1-Measure	0.44	0.83	0.83
GDPR-uncritical	Precision	0.53	0.79	0.78
	Recall	0.96	0.79	0.81
	F1-Measure	0.68	0.79	0.80
Total	Precision	0.73	0.81	0.81
	Recall	0.59	0.81	0.81
	F1-Measure	0.55	0.81	0.81

Tab. 2: Results of the 5-fold cross validation.

The detailed results of our 5-fold cross-validation with a 60/20/20 split in training, validation, and test data are demonstrated in Table 2. It shows the precision, recall, and F1-measure for both classes as well as the total amount of correctly and incorrectly classified tasks. The columns differ in the features used for each model. The left column shows the results when only using the *data item* feature, while the right column demonstrates the results of using all three features in a model. The results show that the discriminating power of the *data item* feature is not very high with an overall F1-measure of 0.55. Still, the precision regarding GDPR-critical tasks is high with a value of 0.90. When only using this *data item* feature, many factors are not taken into account. Some of these factors are considered when using a combination of the *data item* and the *customer relation* features. This addition improves the performance drastically, leading to an F1-measure of 0.81. The addition of the *object* feature to the two prior features leads only to a slight improvement regarding the F1-score of the GDPR-uncritical class. A model using

the *data item*, *customer relation,* and *object* features results in a F1-measure of 0.81. Figure 2 shows the receiver operating characteristic (ROC) curves for the full configuration (*data item*, *customer relation*, *object*). The curves are based on only three data points due to the calculation in Weka but they still provide valuable information about the classifier. More precisely, the curve illustrates the quality of a binary classifier by representing the true positive rate and false positive rate. The curve of a random classifier would result in the point (0.5, 0.5) creating a diagonal from (0, 0) to (1, 1) [Fa06]. The points of the two classes are (0.18, 0.81) for GDPR-critical and (0.19, 0.81) for GDPR-uncritical. This means that our classifier performs satisfactorily.

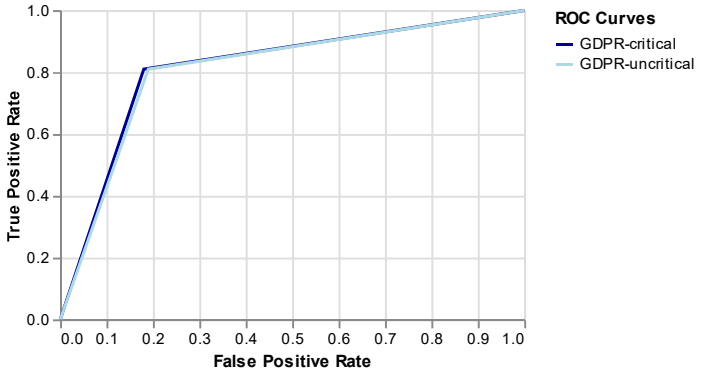

Fig. 2: ROC curves for both classes

When analyzing the errors of the classification, we observe that most errors were caused by the same feature values being used differently. For instance, tasks with the object "information" tend to be misclassified because the term information is vague and is used with different meanings. In one process, the task "enter information" refers to parts and quantities necessary to create a product. In another process, the task "record information" refers to the health status of a patient. The proposed *customer relation* feature can remedy some of these errors as it gives additional insight about the *object* feature "information" since the information that is mentioned in direct contact with a customer or patient is more likely to contain data about the respective customer or patient. Still, there are cases without direct relation to the customer where personal data is dealt with and where ambiguous terms are used. For example, this could be the case when customer data is processed by internal departments after it was obtained from the customer service. Another similar type of error is caused by polysemes, for example in tasks with the *object* value "order". In the task "receive order", the word order refers to data sent by a customer including information about the customer, such as the name, address, and other. In the task "give order" in the context of a hotel, the word order refers to an order to a sommelier to fetch wine. These polysemes seem to be a problem for our approach since they lead to the same feature value but a different GDPR-criticality. Another cause of errors is the lack of training data for

some feature values. Tasks with the *data item* feature value "customer data" were never misclassified since there are several tasks with this feature value. A task with the *data item* feature value "treatment plan" was misclassified although a medical treatment plan will always contain personal data. The problem is that the task "formulate treatment plan" only occurred once in the dataset, which means that it can be misclassified when in the test set. In summary, the first main cause of errors is ambiguity in features leading to feature values with different usage. The second main cause of errors is a lack of training data for some values. It should be noted that Leopold et al. [LvR18] faced similar types of misclassifications in their research. Although these misclassifications exist, we consider the results of the evaluation to be positive. The proposed approach generated satisfactory results in identifying GDPR-critical tasks in textual business process descriptions.

7 Conclusion and Next Steps of Research

The goal of our ongoing research is to create an approach that automatically identifies GDPR-critical tasks in textual business process descriptions. In this paper, we proposed a conceptual approach and conducted a first evaluation. The classification of unseen textual business process descriptions using the proposed features achieved an overall F1-measure of 0.81 meaning that the approach generated satisfactory results.

Although the results of our evaluation were positive, there are limitations to our approach that need to be considered. Firstly, our dataset is not representative. Textual business process descriptions in practice might differ from the ones in our dataset. Depending on the degree of such differences, this could make a correct classification difficult. However, by choosing a dataset that includes different types of descriptions from various sources, we attempted to maximize the validity of our evaluation. Secondly, dealing with tasks from business processes as a source for categorical features poses challenges. Unseen tasks from textual business process descriptions often contain previously unobserved objects and data items since activities and wording in business processes are heterogeneous. As discussed in the previous chapter, this can lead to misclassifications. This can be improved by using other NLP methods, such as word embeddings or language models. Nevertheless, our evaluation has shown that despite these limitations our model can predict the GDPR-criticality of tasks from unseen textual business process descriptions of different sources with satisfactory results. Lastly, in practice, the highest recall of 0.83 for GDPR-critical tasks is not high enough for a fully automatic identification since fines for not complying with the GDPR can be steep. Therefore, our approach cannot be seen as a substitution for a compliance officer when identifying GDPR-critical tasks. Nevertheless, it provides valuable insights and can propose a classification of tasks to support a compliance officer when making the final decision. This means that it can play a role similar to machine learning applications in medicine as described by Forsting [Fo17] for example, where physicians are supported by machine learning models but make the final decision since they are the ones responsible. At the very least, it can give an overview of which business processes often deal with personal data.

During the next steps of our ongoing research, we plan to conduct a second iteration of our DSR project. In this iteration, we will improve the approach by applying a large language model and test it on different datasets. This would improve our model regarding the limitations of the unrepresentative dataset as well as of the heterogeneity of the feature values extracted from business processes. Also, we plan to improve our model by testing other features. After these steps, a summative evaluation of the approach can be made [VPB16]. In future work, we plan to apply similar approaches to different types of sensitive data to generate more insights from textual business process descriptions in this context.

Acknowledgements

The project on which this study is based was funded by the German Federal Ministry of Education and Research under grant number 16KIS1331. The responsibility for the content of this publication lies with the authors.

References

[Ac21] Accenture: State of Cybersecurity Resilience 2021. https://www.accenture.com/content/dam/accenture/final/a-com-migration/custom/us-en/invest-cyber-resilience/pdf/Accenture-State-Of-Cybersecurity-2021.pdf, accessed 8 Jun 2023.

[BRU00] Becker, J.; Rosemann, M.; Uthmann, C. von: Guidelines of Business Process Modeling: Business Process Management. Springer, Berlin, Heidelberg, pp. 30–49, 2000.

[Ca19] Caramujo, J. et al.: RSL-IL4Privacy: a domain-specific language for the rigorous specification of privacy policies. Requirements Engineering 1/24, pp. 1–26, 2019.

[CCR04] Cavusoglu, H.; Cavusoglu, H.; Raghunathan, S.: Economics of IT Security Management: Four Improvements to Current Security Practices. Communications of the Association for Information Systems 14, 2004.

[CV95] Cortes, C.; Vapnik, V.: Support-vector networks. Machine Learning 3/20, pp. 273–297, 1995.

[Fa06] Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters 8/27, pp. 861–874, 2006.

[FMP11] Friedrich, F.; Mendling, J.; Puhlmann, F.: Process model generation from natural language text: Advanced Information Systems Engineering: 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings 23, pp. 482–496, 2011.

[Fo17] Forsting, M.: Machine learning will change medicine. Journal of Nuclear Medicine 3/58, pp. 357–358, 2017.

[GDS19] Ghosh, S.; Dasgupta, A.; Swetapadma, A.: A study on support vector machine based linear and non-linear pattern classification: 2019 International Conference on Intelligent Sustainable Systems (ICISS), pp. 24–28, 2019.

[Ha09] Hall, M. et al.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 1/11, pp. 10–18, 2009.

[HM15] Hirschberg, J.; Manning, C. D.: Advances in natural language processing. Science 6245/349, pp. 261–266, 2015.

[HTF09] Hastie, T.; Tibshirani, R.; Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.

[Jo05] Joachims, T.: Text categorization with support vector machines: Learning with many relevant features: Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21-23, 1998 Proceedings, pp. 137–142, 2005.

[KM03] Klein, D.; Manning, C. D.: Accurate unlexicalized parsing: Proceedings of the 41st annual meeting of the association for computational linguistics, pp. 423–430, 2003.

[KS14] Kumar, N. N.; Shyamasundar, R. K.: Realizing purpose-based privacy policies succinctly via information-flow labels: 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, pp. 753–760, 2014.

[Kü21] Kühnel, S.; Sackmann, S.; Trang, S.; Nastjuk, I.; Matschak, T.; Niezela, L.; Nake, L.: Towards a Business Process-based Economic Evaluation and Selection of IT Security Measures, CEUR Workshop Proceedings 2966, CEUR-WS.org 2021, pp. 7-2, 2021.

[La15] La Rosa, M.: Strategic business process management: Proceedings of the 2015 International Conference on Software and System Process, pp. 177–178, 2015.

[Li11] Liu, B.: Supervised learning. Springer, 2011.

[LvR18] Leopold, H.; van der Aa, H.; Reijers, H. A.: Identifying candidate tasks for robotic process automation in textual process descriptions: Enterprise, Business-Process and Information Systems Modeling: 19th International Conference, BPMDS 2018, 23rd International Conference, EMMSAD 2018, Held at CAiSE 2018, Tallinn, Estonia, June 11-12, 2018, Proceedings 19, pp. 67–81, 2018.

[Pu19] Pullonen, P. et al.: Privacy-enhanced BPMN: enabling data privacy analysis in business processes models. Software and Systems Modeling 18, pp. 3235–3264, 2019.

[Ra12] Ramezani, E. et al.: Separating compliance management and business process management: Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part II 9, pp. 459–464, 2012.

[RLT22] Rahat, T. A.; Long, M.; Tian, Y.: Is Your Policy Compliant? A Deep Learning-based Empirical Study of Privacy Policies' Compliance with GDPR: Proceedings of the 21st Workshop on Privacy in the Electronic Society, pp. 89–102, 2022.

[RMv09] Reijers, H. A.; Mans, R. S.; van der Toorn, R. A.: Improved model management with aggregated business process models. Data & Knowledge Engineering 2/68, pp. 221–243, 2009.

[RRM21] Rizun, N.; Revina, A.; Meister, V. G.: Analyzing content of tasks in Business Process Management. Blending task execution and organization perspectives. Computers in Industry 130, p. 103463, 2021.

[Ru10] Russell, S. J.: Artificial intelligence a modern approach. Pearson Education, Inc, 2010.

[Sc05] Schuler, K. K.: VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania, 2005.

[Se20] Serrado, J. et al.: Information security frameworks for assisting GDPR compliance in banking industry. Digital Policy, Regulation and Governance 3/22, pp. 227–244, 2020.

[SKS18] Sackmann, S.; Kuehnel, S.; Seyffarth, T.: Using business process compliance approaches for compliance management with regard to digitization: evidence from a systematic literature review: Business Process Management: 16th International Conference, BPM 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings 16, pp. 409–425, 2018.

[TK01] Tong, S.; Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research Nov/2, pp. 45–66, 2001.

[To20] Torre, D. et al.: An ai-assisted approach for checking the completeness of privacy policies against gdpr: 2020 IEEE 28th International Requirements Engineering Conference (RE), pp. 136–146, 2020.

[VK15] Vaishnavi, V. K.; Kuechler, W.: Design science research methods and patterns: innovating information and communication technology. Crc Press, 2015.

[vLR17] van der Aa, H.; Leopold, H.; Reijers, H. A.: Comparing textual descriptions to process models-the automatic detection of inconsistencies. Information Systems 64, pp. 447–460, 2017.

[VPB16] Venable, J.; Pries-Heje, J.; Baskerville, R.: FEDS: a framework for evaluation in design science research. European journal of information systems 25, pp. 77–89, 2016.

[We19] Weske, M.: Business process management. Concepts, languages, architectures. Springer, Berlin, Heidelberg, 2019.