Towards a Privacy Compliant Cloud Architecture for Natural Language Processing Platforms

Matthias Blohm[1], Claudia Dukino[2], Maximilien Kintz[2], Monika Kochanowski[2], Falko Koetter[2] and Thomas Renner[2]

[1] *University of Stuttgart IAT, Institute of Human Factors and Technology Management, Germany*

[2] *Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Germany*

Keywords: Natural Language Processing, Artificial Intelligence, Cloud Platform, GDPR, Compliance, Anonymization.

Abstract: Natural language processing in combination with advances in artificial intelligence is on the rise. However, compliance constraints while handling personal data in many types of documents hinder various application scenarios. We describe the challenges of working with personal and particularly sensitive data in practice with three different use cases. We present the anonymization bootstrap challenge in creating a prototype in a cloud environment. Finally, we outline an architecture for privacy compliant AI cloud applications and an anonymization tool. With these preliminary results, we describe future work in bridging privacy and AI.

1 INTRODUCTION

Natural language processing (NLP) is on its rise. Researchers all over the scientific landscape investigate manifold real world applications. However, in these application scenarios the General Data Protection Regulation (European Union, 2016) is conceived as a major challenge in NLP. This is, because in contrast to tabular data, anonymization by aggregation is not possible for natural language text, as shown in Figure 1. Furthermore, pseudonymization methods can cause information loss.

These issues are all the more crucial when cloud-based solutions are considered. In order to make automated text analysis widely available, to share knowledge across stakeholders and to reduce tagging workload, cloud-based text analysis platforms are a promising solution. However, working with GDPR-relevant data in the cloud is particularly difficult. Thus, the need for ways of taking advantages of cloud solutions while remaining GDPR-compliant increases.

A solution for automatically dealing with GDPR relevant data especially in natural language documents is often missing. Therefore, anonymization and pseudonymization is done manually. A promising idea is to use artificial intelligence (AI) / machine learning (ML) for anonymizing natural language documents - however, to train this artificial intelligence, non-anonymized and anonymized documents are needed. To get around this problem, several options are possible.

This paper is structured as follows. Section 2 describes related work on the topics of natural language processing, anonymization and pseudonymization as well as platforms. Section 3 describes three existing application scenarios - court decisions, healthcare and insurance fraud. Based on these application scenarios, a central research question is derived in Section 4. To answer this question, section 5 outlines a solution architecture for GDPR-compliant, semi-automated document anonymization as well as an in-progress prototype. Finally, Section 6 summarizes the work and gives an outlook on research-in-progress.

2 RELATED WORK

We describe related work in three areas: (1) NLP in GDPR context and (2) anonymization and pseudonymization by artificial intelligence as well as (3) platform solutions for NLP.

(1) Currently the possible slowdown of Europe's innovation progress especially in the field of Text and Data Mining (TDM) due to restrictive laws of data protection and privacy is an important issue in public discussions (European Comission, 2014). Since

Blohm, M., Dukino, C., Kintz, M., Kochanowski, M., Koetter, F. and Renner, T.
Towards a Privacy Compliant Cloud Architecture for Natural Language Processing Platforms.
DOI: 10.5220/0007746204540461
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 454-461
ISBN: 978-989-758-372-8

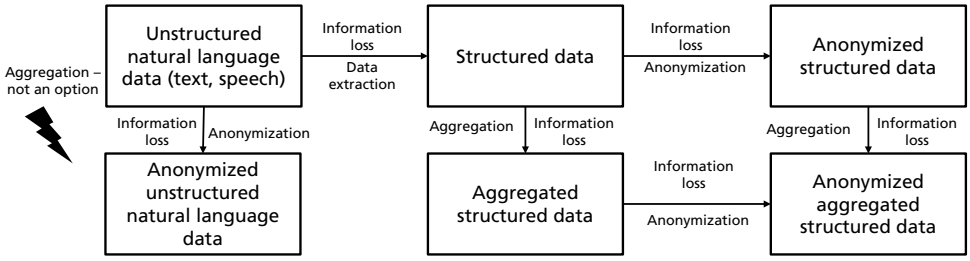

Figure 1: Natural language data cannot be anonymized by aggregation before working with it, as it is done e.g. with tabular data in sensitive contexts.

the introduction of the GDPR in Europe, some considerations have been made about its compliance with purposes of artificial intelligence. On one hand, new restrictions of data privacy indeed complicate the data acquisition for machine learning tasks. On the other hand, data protection laws may also encourage a fairer and more transparent processing of personal data (Kamarinou et al., 2016). Popular software that was trained with the means of machine learning to automatically identify and protect sensitive personal data is for example given with *Amazon Macie* or *Google DLP* (Marko, 2017).

(2) The importance of novel anonymization and pseudonymization techniques is underlined by proclaimed challenges such as the NLP Shared Tasks announced by i2b2, where often one of the tasks was to de-identify personal data in clinical reports (i2b2 Informatics for Integrating Biology & the Bedside, 2019). In 2014's challenge the University of Nottingham achieved the highest f1-score of 93.6% correctly recognized entities by combining machine learning and rule-based techniques (Yang and Garibaldi, 2015). For tackling the problem of de-identification, a common way is to rely on named entity recognition (NER) for detecting sensitive information even in larger unstructured text documents (Vincze and Farkas, 2014). A promising approach could be to combine the results of several different entity recognizers with coreference resolution processing in order to find and replace a maximum of entities such as proper names, places or dates, while maintaining full meaning in the document context (Dias, 2016).

(3) In the business domain, countless scenarios are thinkable in which companies could benefit from using AI, for example for supporting classification and decision tasks as automatic customer claim handling (Coussement and den Poel, 2008; Yang et al., 2018). Nowadays several providers like *Aylien* (AYLIEN, 2019) already offer AI platforms for natural language processing as a self-service. Here customers can build and train individual models for textual processing without the need of any programming skills. However, sending sensitive data to cloud servers is still a critical issue to deal with when using those platforms. Therefore, some providers like *Lexalytics* (Lexalytics, 2019) also offer on premise solutions of their software which can be installed and run only locally on internal hardware.

Altogether, machine learning has been shown as applicable for improving anonymization or pseudonymization. Many state of the art approaches exist therefore. However, for being able to process text documents in a cloud environment, a practicable solution for training these algorithms without the need of on premise solutions in a multi-party environment is necessary. To the best of our knowledge, a complete solution for this task has not yet been described. We formulate the research question in Section 4.

3 APPLICATION SCENARIOS

We describe three application scenarios: court decisions, healthcare and fraud detection, having in common: (1) personal data is included all of the time, (2) particularly sensitive data is included often, and (3) high potential for machine learning in textual documents is given.

3.1 Court Decisions

In Germany, court decisions generally have to be made available to the public upon request. However, to protect the privacy of the parties involved, the judicial decisions must be anonymized prior to publication. Especially in criminal and family law, court decisions often contain sensitive data, e.g. the biography of the accused, or private details of family life.

Important court decisions are published by the courts on their own accords. Other court decisions are requested for an administrative fee. While case law is generally not as important in Germany as in other jurisdictions like the USA, requests for court decisions are increasing.

Currently, court decisions are anonymized manually by judges or clerks, resulting in a considerable time investment for these highly skilled workers, which could be used elsewhere.

3.2 Healthcare

The healthcare sector is one of the most highly regulated sectors in respect to data protection, as most documents contain sensitive data of patients.

The healthcare sector is under pressure by rising healthcare costs, an aging populace, a shortage of physicians, as well as comprehensive documentation requirements (Meinel and Koppenhagen, 2015). While IT is widely used in areas like diagnostics and robotics, adoption of cloud applications is slow (Lux et al., 2017). One reason for this is the challenge to comply with data protection laws. On the other hand, many scenarios could profit from sharing anonymized documents, ranging from standard services like translation of medical instructions into a patient's native language to cooperative diagnosis and medical research.

One challenge in the healthcare area is that not only directly identifying data (e.g. name, address), but also indirectly identifying data (e.g. combination of symptoms, rare diseases) has to be removed. Determining what data is indirectly identifying requires expert expertise. How and if such a determination could be performed automatically is an open research question.

3.3 Fraud Detection

Undetected insurance fraud costs insurers billions of dollars every year (Power and Power, 2015). To counteract these losses, insurance companies try to detect fraud before payments are made. Conventional fraud detection relies on manual work as well as IT solutions, which perform a rule-based analysis on a claim. These rules are created and maintained by domain experts and focus on structured data that is known about the claim. Unstructured documents and images are typically investigated manually.

As a decision problem fraud detection could possibly be improved by applying ML. Depending on the type of insurance, fraud rates are claimed by insurance companies to be as high as 50 percent (smartphone insurance). However, it can be assumed that not all fraudulent claims are detected as such. For example, a claim may be abandoned by a claimant if additional questions are asked, making it unclear why no payment took place.

Insurance companies could improve fraud detection by sharing anonymized claim data in order to build a communal AI (Power and Power, 2015). Data protection laws necessitate anonymization or pseudonymization of this data. This concerns not only personal data, but also image files (e.g. license plates on damaged cars).

4 RESEARCH QUESTION

Cloud computing, big data and artificial intelligence make many new application scenarios possible. The current public dialogue about artificial intelligence and the digital transformation have made many organizations aware of these new possibilities (IDC, 2018). On the other hand, organizations have been sensitized to privacy concerns by the public dialogue about the GDPR.

This creates a perceived conflict between technical possibilities and legal requirements. In our work with organizations in research and industry projects we found data protection concerns to be the greatest perceived challenge to overcome. In an ongoing Fraunhofer survey of over 200 German organizations, data protection was named the greatest challenge when using AI [1]. As new machine learning algorithms are tailored for large amounts of data, questions of data protection need to be solved before building even an exploratory prototype. For tabular data, aggregating data for ensuring privacy may be an option. However, for textual data this is not possible. It is possible to extract structured data from text and then apply machine learning - however, text processing relies on more information than just the structured contents of the text documents. Therefore, the information loss by working in this way is not acceptable for most natural language processing machine learning scenarios. It is therefore necessary to work with the original documents and to anonymize or pseudonymize these.

Additionally, finding the personal and particularly sensitive data is a challenge. Table 1 shows how good various state-of-the-art methods work for identifying text with known and unknown formats and value sets. If the format is known, like for example license plates, finding the entity and anonymization is easier than if the format is unknown. Dates of birth can be for example contained in various forms in a document. If the format is known, like an e-mail-address, it is easy

[1] At the time of review, this survey is still open for participation. In the camera ready paper, we will update this sentence with the final results. The survey can be found at: https://www.befragung.iao.fraunhofer.de/index.php/568823

Table 1: Methods for anonymizing data in text documents.

method **difficulty** *examples*	Format Known	Format Unknown
Valueset Known	reference list **very easy** *invoiceId...*	AI / ML **medium** *birth date, priority,...*
Valueset Unknown	rules, regular expressions **easy** *e-mail, IBAN,...*	AI / ML **hard** *political opinion,...*

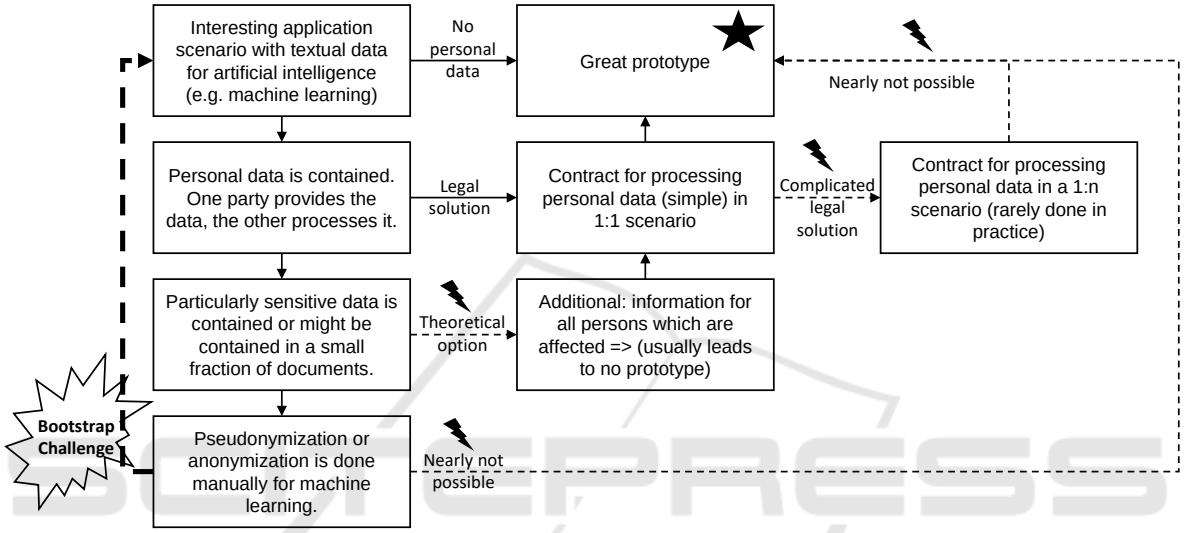

Figure 2: Alternatives and possible outcomes for prototype development with GDPR relevant data.

to identify and anonymize data, even with unknown datasets. This is extremely useful for personal data like IBANs and social security numbers. However, in the field of particularly sensitive data, like ethnical background, sexual orientation, or political affiliation, the format as well as the values are unknown. This makes finding this information a very difficult task, making it hard to anonymize as well. Additionally, machine learning algorithms need lots of data to handle this kind of difficult questions. Finally, this makes machine learning algorithms the most promising approach for solving this problem.

However, solving this in machine learning gives the need for a prototype. Figure 2 shows possible approaches for implementing prototypes while remaining GDPR compliant.

A solution for data processing is possible, as long as consent was received by the data provider, no sensitive data is contained, and a contract for data processing taking into account GDPR is made.

Special care needs to be taken when sensitive data (e.g. health data, information about racial or ethnical background) might be contained in the documents to be used. While the GDPR allows an exception for using sensitive data in research, processing it requires explicit consent of all affected persons. As projects with real companies rely on a large, existing volume of data, this is not feasible, as all affected customers would need to explicitly give consent.

As an additional challenge, most organizations, especially small and medium enterprises, lack the skills and data volume to realize AI projects in-house, so they are dependent on service providers or cooperation to pool data. In this case, a contract between multiple parties would be necessary, complicating a possible legal solution and making GDPR compliance questionable.

As an alternative, the GDPR allows the anonymization of documents. Once a dataset has been anonymized and individuals are no longer identifiable, GDPR no longer applies. The solution in existing projects was manual tagging and anonymization of large volumes of documents. This work needs to be performed in-house, as outsourcing it would present a compliance violation as well. While this is possible if companies receive a research grant,

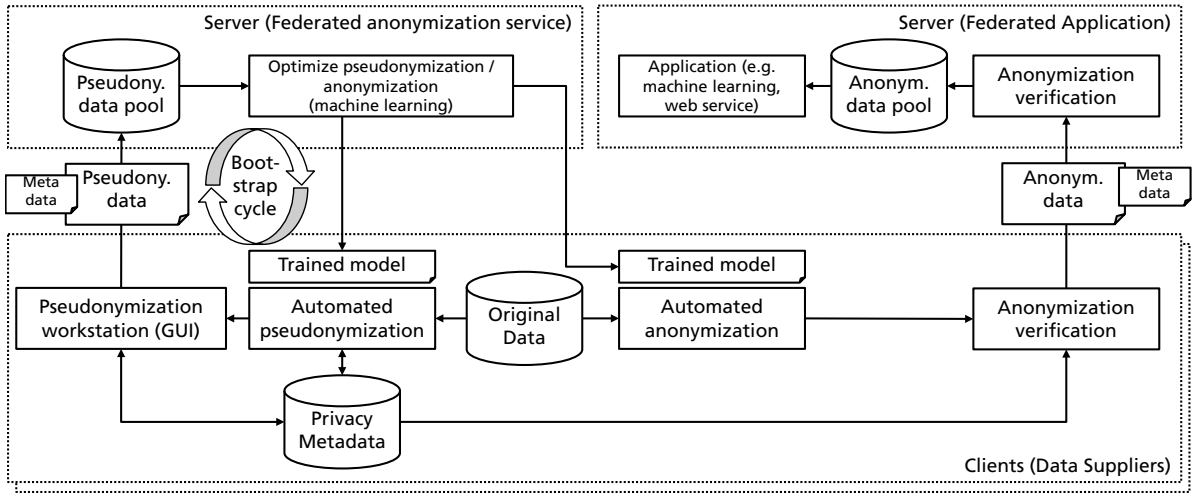

Figure 3: Architecture outline for federated pseudonymization and anonymization for NLP and AI. Private data remains on data supplier's systems. A trained model for pseudonymization and anonymization is iteratively developed in a bootstrap cycle. Better and better automation is used to pseudonymize larger and larger amounts of private data, resulting in a final model used for anonymization. Anonymization is verified with the private data tagged during pseudonymization.

generally it is not economically feasible.

The alternatives shown in Figure 2 represent a challenge for applied research into AI and NLP. In preliminary talks with companies interested in research participation, we found the manual effort for anonymization on the scope AI requires to be a dealbreaker.

Thus, we formulated a preliminary **research question**: To conduct our AI research, we need to develop tools and methods to aid in anonymization of documents. Possible approaches to anonymization that come to mind are of course AI and NLP. This creates a **bootstrap challenge**, as to optimize and customize anonymization methods for a certain class of documents, access to these documents is necessary.

How can this bootstrap challenge be solved within a cloud environment with software tools to lessen manual effort for anonymization while maintaining compliance with GDPR?

5 ARCHITECTURE AND PROTOTYPE

To solve the *bootstrap challenge*, we outlined a software solution for federated data pseudonymization and anonymization, of which the architecture is shown in Figure 3.

This solution uses iterative pseudonymization of documents in order to train a domain-specific model for anonymization. The reason to use pseudonymization first is to provide pseudo-non-anonymized documents to a service provider without giving the provider real documents.

Pseudonymization is performed both automatically and manually. In a first step, a small sample of available documents are pseudonymized with a generic pseudonymization algorithm, pseudonymizing common data items like e-mail addresses, names and phone numbers. This small sample is manually reviewed on a pseudonymization workstation, which allows correcting and amending the automated pseudonymization, e.g. by tagging personal data that was not automatically pseudonimized and by detagging false positives.

The result of this pseudonymization is a set of pseudonymized documents as well as a set of privacy metadata (i.e. private data items and their positions in documents). The pseudonymized documents as well as reduced privacy metadata (indicating position, but not private data) are sent to a federated anonymization service. If multiple organizations want to pool their data, this step is performed independently by every data provider.

Based on the received pseudonymized data, the anonymization service can improve the generic model, either using machine learning or by manually extending the model. The methods for extraction types of data have been discussed in Section 4 and are shown in an overview in Table 1.

This process is repeated with the improved model. This time, a different, possibly sample of documents is used. If the improvement was successful, the manual pseudonymization effort should be reduced. With the additional data, the trained model can be further improved. Iteration in this *bootstrap cycle* continues

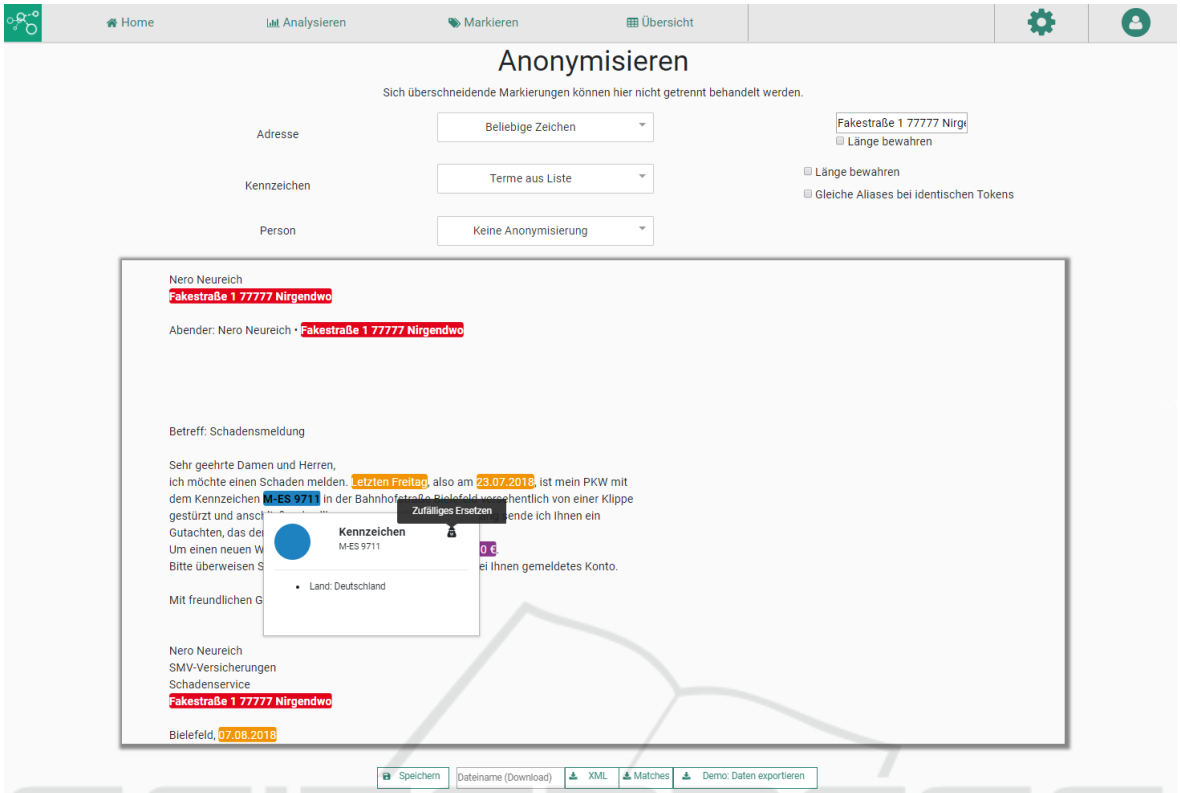

Figure 4: Screenshot of entity recognition, automatic anonymization and manual correction workbench (working name Textominado). The software allows to automatically detect several kinds of entities, for example names, addresses, license plates, and to replace them with pseudonymes or just removing them preserving length or not. Focus is the interaction of automatic detection and anonymization as well as manual corrections, which can be used for machine learning in the next step.

until a model of acceptable quality is obtained.

With this model, anonymization can be performed. Compared to pseudonymization, private data is not substituted by equivalent data, but excised. Anonymization is performed on the whole set of original data, and is verified automatically using the privacy metadata obtained during pseudonymization. Depending on the sensitivity, random spot checks could be performed as well.

The anonymized data may then be passed to an application provider for use in the actual business case (e.g. court decision database, medical research, fraud detection). The application provider can use the same anonymization verification component to exercise due diligence after receiving data.

To aid in future data acquisition in our research projects, we plan to fully implement this software solution. As an immediate solution, we partially implemented this approach. The platform Textominado allows manual and extensible automatic pseudonymization/anonymization to create data for machine learning that is compliant with the GDPR.

Textominado's flexible architecture consists of a *Java Spring* backend with loosely coupled modules that allows quickly adding and modifying of service endpoints. This way, we are able to integrate any kind of library for providing **analysis functions** like entity recognizers that help locate sensitive personal data in unstructured text documents. New endpoints are registered automatically in the frontend with their corresponding URL and can directly be applied to an input text together with any other tool provided this way. Furthermore, the UI that we built using *nodejs* and *react* supports easy **manual tagging** of additional crucial entities, which may have not been detected by the analysis tools. Finally, with the built-in **pseudonymization/anonymization** feature, which is shown in Figure 4 we are able to create different kinds of meaningful pseudonyms for each category of tagged entity. We are also experimenting with coreference resolution in order to prevent a greater loss of information by keeping the replacement of words consistent throughout the whole textual contents.

With the introduction of Textominado we took a first step towards offering a powerful platform that facilitates customized creation of anonymized data for

machine learning purposes. The pseudonymization can be improved and extended by implementing custom libraries, but this is a manual process. Our goal is to automate improvement processes of entity recognition and anonymization by including self-learning components, that are able to improve the automated detections based on the users' manual corrections on the results of the analysis tools, thus realizing the *bootstrapping cycle* outlined above.

6 CONCLUSION AND OUTLOOK

In this work we have described the challenges of performing applied research in NLP and AI with real unstructured data while remaining compliant with the GDPR and other data protection laws.

We have shown the need to anonymize and share data in three application areas: court decision, healthcare and insurance fraud detection. From practical experience in research projects, we have outlined challenges and possible solutions for obtaining data to develop research prototypes. Based on these experiences, we have defined a *bootstrap challenge*: AI and NLP can be used to automate data anonymization for research, but anonymized data is needed to create AI and NLP anonymization solutions in the first place. The resulting research questions is how to solve this bootstrap problem while lessening manual effort for anonymization.

We have outlined a possible solution architecture, which incrementally improves domain-specific pseudonymization in a *bootstrap cycle*, thus solving the bootstrap challenge, and shown *Textominado*, a prototype for pseudonymization and anonymization of unstructured documents.

Since the contents discussed in this paper are still ongoing research, no evaluation has been done for our prototype yet. But talking to different companies and public organizations revealed that there is indeed a big need for practicable ways of anonymizing unstructured textual data. In future research, we plan to use *Textominado* to acquire anonymized data from real-world organizations for use in AI and NLP research projects. In this process, we plan to extend *Textominado* in order to implement the outlined solution architecture and investigate its feasibility.

ACKNOWLEDGEMENTS

This work was partly supported by the project SmartAIwork, which is funded by the Federal Ministry of Education and Research (BMBF) under the funding number 02L17B00ff. We like to thank our contacts at the court and in the healthcare and insurance industry as well as the students working in student projects for their efforts.

REFERENCES

AYLIEN (2019). Text analysis platform — custom nlp models. https://aylien.com/text-analysis-platform/.

Coussement, K. and den Poel, D. V. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4):870 – 882.

Dias, F. M. C. (2016). Multilingual automated text anonymization. Master's thesis, Instituto Superior Técnico, Lisboa.

European Comission (2014). Text and data mining - report from the expert group.

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

i2b2 Informatics for Integrating Biology & the Bedside (2019). 2016 cegs n-grid shared-tasks and workshop on challenges in natural language processing for clinical data. https://www.i2b2.org/NLP/.

IDC (2018). Multi-Client-Studie Künstliche Intelligenz und Machine Learning in Deutschland 2018. https://idc.de/de/research/multi-client-projekte/kunstliche-intelligenz-und-machine-learning-in-deutschland-die-nachste-stufe-der-datenrevolution/kunstliche-intelligenz-und-machine-learning-in-deutschland-projektergebnisse.

Kamarinou, D., Millard, C., and Singh, J. (2016). Machine learning with personal data. In *Queen Mary School of Law Legal Studies Research Paper No. 247/2016*. SSRN.

Lexalytics (2019). Salience 6, lexalytics state of the art natural language processing engine on your own hardware. https://www.lexalytics.com/salience/server.

Lux, T., Breil, B., Dörries, M., Gensorowsky, D., Greiner, W., Pfeiffer, D., Rebitschek, F. G., Gigerenzer, G., and Wagner, G. G. (2017). Healthcare — between privacy and state-of-the-art medical technology. *Wirtschaftsdienst*, 97(10).

Marko, K. (2017). Using machine intelligence to protect sensitive data. https://diginomica.com/2017/08/24/using-machine-intelligence-protect-sensitive-data/.

Meinel, C. and Koppenhagen, N. (2015). Thesenpapier zum Schwerpunktthema Smart Data im Gesundheitswesen (in German). https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/Smart_Data_Thesenpapier_SmartData_Gesundheitswesen.html.

Power, D. J. and Power, M. L. (2015). Sharing and analyzing data to reduce insurance fraud. In *MWAIS 2015 Proceedings*.

Vincze, V. and Farkas, R. (2014). De-identification in natural language processing. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1300–1303.

Yang, H. and Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30 – S38. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Yang, Y., Xu, D.-L., Yang, J.-B., and Chen, Y.-W. (2018). An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications. *Knowledge-Based Systems*, 162:202 – 210. Special Issue on intelligent decision-making and consensus under uncertainty in inconsistent and dynamic environments.