UNIVERSIDADE D
COIMBRA

Paulo Miguel Guimarães da Silva

CONTRIBUTIONS TO PERSONAL DATA PROTECTION AND PRIVACY PRESERVATION IN CLOUD ENVIRONMENTS

Tese no âmbito do Programa de Doutoramento em Ciências e Tecnologias da Informação, orientada pelo Professor Doutor Edmundo Monteiro, e pelo Professor Doutor Paulo Simões, e apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Dezembro de 2020

UNIVERSIDADE D
COIMBRA

Department of Informatics Engineering
Faculty of Sciences and Technology
University of Coimbra

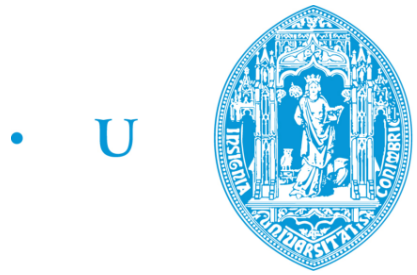
Contributions to Personal Data Protection and Privacy Preservation in Cloud Environments

Paulo Miguel Guimarães da Silva

Doctoral Program in Information Science and Technology
PhD Thesis submitted to the University of Coimbra

Advised by Prof. Dr. Edmundo Monteiro and Prof. Dr. Paulo Simões

December, 2020



Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Contribuições para a proteção de dados pessoais e preservação da privacidade em ambientes de nuvem

Paulo Miguel Guimarães da Silva

Programa de Doutoramento em Ciências e Tecnologias da Informação
Tese de Doutoramento apresentada à Universidade de Coimbra

Orientado pelo Prof. Dr. Edmundo Monteiro e pelo Prof. Dr. Paulo Simões

Dezembro, 2020

This work was partially supported by the European Union (EU) H2020 and the Brazilian (BR) Ministry of Science, Technology, Innovation, and Communications/National Research Network, through the projects EUBra–BIGSEA (BR 23614/EU 690116) and ATMOSPHERE (BR 51119/EU 777154); and by the PoSeID-on project - Protection and control of Secured Information by means of a privacy enhanced Dashboard, Grant Agreement Number: 786713, H2020-DS-2016-2017/ DS-08-2017.

Acknowledgements

A project like this thesis is hardly ever the work of anyone alone. The contributions of many different people, in different ways and various places, have made this possible. I want to extend my appreciation, especially to the following.

Firstly, I would like to thank my advisors, Edmundo Monteiro and Paulo Simões. The interest and availability demonstrated from the very first moment was a crucial motivation. Their engagement was felt the whole time, and I know I can count on them now and in the times to come. For all the help, dedication, guidance and expertise they provided me, academically and personally, I truly owe them all my appreciation. Without them, this would not be possible.

I would also like to express my gratitude towards the rest of the team I was working with on the PoSeID-on project: Marilia Curado, Nuno Antunes, Fernando Boavida, Rui Casaleiro, Carolina Gonçalves and Carlos Junior. Their kind help and support have made this journey much more exciting.

In addition, I would like to thank the Department of Informatics Engineering (DEI), the Centre for Informatics and Systems of the University of Coimbra (CISUC), and of course the Laboratory of Communications and Telematics (LCT).

The days would have passed far more slowly without the support of my friends and colleagues. They helped me overcome difficulties and maintain focus on my work.

I wish to thank my family for the love and support they provided me through my life, without whom I would never have enjoyed so many opportunities. I must equally acknowledge my wife Simona, whose love, encouragement and patience, helped make this challenging journey a little bit easier to accomplish.

Abstract

Personal data is currently being used in countless applications in a vast number of areas. Despite national and international legislation, the fact is that individuals still have little to no control over who uses their data and for what purposes. As regulations vary from region to region, data is often stored and processed in multiple locations by multiple data processors. Moreover, the security concerns of a system are sometimes addressed individually or in an ad-hoc manner, which may result in inadequate solutions. In the end, data protection and privacy assurances are still, in many cases, only a theoretical possibility. As such, it is necessary to propose mechanisms that maximise data protection and provide increased privacy assurances.

A strategy to ensure appropriate levels of security and privacy is mandatory. In this work, it was possible to design, develop and evaluate mechanisms that fill the issues mentioned above. One of the pillars of this strategy is the inclusion of Authentication, Authorisation and Accounting (AAA) solutions that securely control access to individuals' data. The other pillar relies on the usage of intelligent, automated, and non-intrusive mechanisms that monitor and control personal data to increase privacy assurances.

To fulfil such strategy, the development of a cloud-based AAA solution was the very first step to control individuals' access to data. The proposed solution is composed of a reverse proxy, a custom web application and a NoSQL database.

The mechanisms proposed in this thesis recur to Natural Language Processing (NLP), Named Entity Recognition (NER) and Machine Learning (ML) algorithms in a hybrid approach. A series of NER models capable of identifying personal information are also trained with algorithms such as Multi-Layer Perceptron (MLP) and Random Forests (RF), using only publicly available datasets as a source of training and validation data.

The mechanisms proposed in this work comply with existing regulations and are designed under appropriate cloud-based deployment and life cycle management strategies. Moreover, this thesis proposes a fuzzy privacy risk model that allows the assessment of privacy risk levels associated with data transactions.

The advantages and drawbacks of the proposed mechanisms were evaluated in pilot use cases in the scope of two international projects: H2020 EUBra-BIGSEA and H2020 PoSeID-on. The evaluation conducted on both technical and user-centred scenarios indicates that the proposed mechanisms have high data classifying accuracy, support large volumes of data with distinct characteristics and to increase individuals' privacy awareness and control.

Keywords: Privacy Enhancing Technologies; Privacy Risk Assessment; Cloud Systems; Personally Identifiable Information; Machine Learning; Natural Language Processing

Resumo

Os dados pessoais são atualmente utilizados em inúmeras aplicações num grande número de áreas. Apesar da legislação nacional e internacional, o facto é que indivíduos ainda têm pouco ou nenhum controlo sobre quem usa os seus dados pessoais, e para que fins. Como os regulamentos variam de região para região, os dados geralmente são armazenados e processados em vários locais, e por vários processadores de dados. Além disso, as questões de segurança dos sistemas por vezes são tratadas individualmente ou de maneira ad-hoc, o que pode resultar em soluções inadequadas. No final, a proteção de dados e as garantias de privacidade ainda são, em muitos casos, apenas uma possibilidade teórica. Como tal, é necessário propor mecanismos que maximizem a proteção de dados e forneçam maiores garantias de privacidade.

Uma estratégia para garantir níveis adequados de segurança e privacidade é obrigatória. Neste trabalho, foi possível projetar, desenvolver e avaliar mecanismos que atendem às questões mencionadas acima. Um dos pilares desta estratégia é a inclusão de soluções de Autenticação, Autorização e Auditabilidade (AAA) que controlam o acesso aos dados pessoais com segurança. O outro pilar depende do uso de mecanismos inteligentes, automatizados e não intrusivos que monitoram e controlam os dados pessoais de modo a aumentar as garantias de privacidade.

Para seguir essa estratégia, o primeiro passo foi o desenvolvimento de uma solução AAA baseada na nuvem, que controla o acesso a dados pessoais. A solução proposta é composta por um procurador reverso, uma aplicação web personalizada e uma base de dados NoSQL.

Os mecanismos propostos nesta tese recorrem a Processamento de Linguagem Natural (PNL), Reconhecimento de Entidades Mencionadas (REM) e Aprendizagem Automática (AA) de uma forma híbrida. Uma série de modelos REM capazes de identificar informações pessoais também são treinados com algoritmos tais como Perceptron Multicamada (PM) e Florestas de Decisão Aleatórias (FDA), usando apenas conjuntos de dados publicamente disponíveis, como fonte de dados de treino e validação.

Os mecanismos propostos neste trabalho estão em conformidade com os regulamentos existentes e são projetados de acordo com uma implementação baseada em nuvem e estratégias de gestão de ciclo de vida apropriadas. Além disso, esta tese propõe um modelo fuzzy de risco de privacidade que permite avaliar os níveis de risco de privacidade associados às transações de dados.

As vantagens e desvantagens dos mecanismos propostos foram avaliadas em casos de uso piloto no âmbito de dois projetos internacionais: H2020 EUBra-BIGSEA e H2020 PoSeID-on. A avaliação realizada em cenários técnicos e centrados no usuário indica que os mecanismos propostos têm alta precisão de classificação de dados, suportam grandes volumes de dados com características distintas e aumentam a perceção e o controle da privacidade dos indivíduos.

Palavras-chave: Tecnologias que Aumentam a Privacidade; Avaliação de Risco de Privacidade; Sistemas em Nuvem; Informação Pessoalmente Identificável; Aprendizagem Automática; Processamento de Linguagem Natural

Foreword

The work described in this thesis was accomplished at the Laboratory of Communication and Telematics (LCT) of the Centre for Informatics and Systems of the University of Coimbra (CISUC), within the context of the following projects:

Project H2020 EUBra-BIGSEA EUBra-BIGSEA is a project funded in the third coordinated call Europe - Brazil focused on the development of advanced Quality of Service (QoS) for Big Data applications, demonstrated in the scope of the Massive Connected Societies. EUBra-BIGSEA developed a framework, a platform and a library to ease the development of highly-scalable, privacy-aware data analytic applications running on top of Quality of Service cloud infrastructures, reducing development cycles and deployment costs.

The results described in this thesis have contributed to the state of the art of EUBra-BIGSEA's security and privacy solutions. Those contributions include a literature review of security and privacy solutions, the definition of the architecture of its security module Authentication, Authorisation and Accounting as a Service (AAAaaS)) as well as its development, integration and validation.

EUBra-BIGSEA was funded by the European Commission under the Cooperation Programme and Horizon 2020 Programme under Grant Agreement Nº 690116.

Project H2020 PoSeID-on PoSeID-on delivered an innovative and intrinsically scalable platform, as a comprehensive solution aimed to safeguard the rights of data subjects, exploiting the cutting-edge technologies of Smart Contracts and Blockchain, as well as support organisations in data management and processing while ensuring General Data Protection Regulation (GDPR) compliance.

The results described in this thesis have contributed to the state of the art of PoSeID-on's privacy solutions. Such contributions include a literature review of privacy and data analysis solutions, an active role in the definition of the whole system architecture and its requirements; the definition of the architecture of the Personal Data Analyser (PDA) module, as well as its development, integration and validation.

PoSeID-on was funded by the European Commission's Horizon 2020 Programme under Grant Agreement Nº 786713.

The outcome of the research, design, experiments, and assessments of several mechanisms on the course of this work resulted in the following publications:

Journal papers:

- Silva, P. et al. (2020). Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges. IEEE Access, Vol. 9, pp. 10473-10497, Print ISSN: 2169-3536. Online ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3049599, January 2021.

  Main contributions: Literature review and analysis of Privacy Enhancing Technologies and respective Cloud Applicability.

- Silva, P. et al. (2020). Risk Management and Privacy Violation Detection in the PoSeID-on Data Privacy Platform. SN COMPUT. SCI. 1, 188. DOI: 10.1007/s42979-020-00198-9.

  Main contributions: collaboration in the specification of the PoSeID-on platform architecture, communication protocol and, privacy violation detection module development and integration.

- Casaleiro, R. et al. (2020). Protection and Control of Personally Identifiable Information: The PoSeID-on Approach. Journal of Data Protection & Privacy, 3(2).

  Main contributions: collaboration on the definition of the PoSeID-on architecture as well as privacy violation detection and risk analysis modules.

- Silva, P. et al. (2018) A Europe-Brazil Context for Secure Data Analytics in the Cloud. IEEE Security & Privacy, vol. 16, no. 6, pp. 52-60. DOI: 10.1109/MSEC.2018.2875326.

  Main contributions: presentation of the BIGSEA platform architecture, Authentication, Authorisation and Accounting (AAA) module and literature review of privacy solutions for the Cloud.

Conference Papers

- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2020). Using NLP and Machine Learning to Detect Data Privacy Violations. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, pp. 972-977. DOI: 10.1109/INFOCOMWKSHPS50562.2020.9162683.

  Main contributions: specification of Natural Language Processing (NLP) pipeline to analyse and detect Personally Identifiable Information (PII) on data exchanges. Training and validation of Named Entity Recognition (NER) models using Multi-Layer Perceptron (MLP) and Random Forest (RF) algorithms. Presentation of the work at the conference.

- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2020). Using Natural Language Processing to Detect Privacy Violations in Online Contracts. In Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC '20). Association for Computing Machinery, New York, NY, USA, 1305–1307. DOI: 10.1145/3341105.3375774.

  Main contributions: analysis of the capabilities of NLP tools (performance and F1-Score) in classifying PII by using general-purpose datasets. Presentation of the work in the conference.

- Silva P., Kencl L. and Monteiro E. (2018). Data Privacy Protection - Concealing Text and Audio with a DNA-inspired Algorithm. In 12th International Conference on Autonomous Infrastructure, Management and Security, AIMS 2018, Munich, Germany, June 4-5, Proceedings (pp. 46–59).

  Main contributions: experimental work with a DNA-inspired data concealing algorithm. Assessment of the algorithm's capabilities, applicability potential. Using the algorithm with text and audio files and generating concealed documents with similar utility. Presentation of the work at the conference.

Other Publications

- Bagnato A., Silva P., Alaqra A.S. and Ermis O. (2020). Workshop on Privacy Challenges in Public and Private Organisations. In: Friedewald M., Önen M., Lievens E., Krenn S., Fricker S. (eds) Privacy and Identity Management. Data for Better Living: AI and Privacy. Privacy and Identity 2019. IFIP Advances in Information and Communication Technology, vol 576. Springer, Cham. DOI: 10.1007/978-3-030-42504-3_6.

  Main contributions: Organisation of workshop. Analysis and reporting of the feedback collected from a questionnaire given to the audience. Presentation of the PoSeID-on platform and its risk and privacy management modules.

- Alic A. et al. (2019). BIGSEA: A Big Data Analytics Platform for Public Transportation Information. Future Generation Computer Systems, 96, 243 - 269. DOI: 10.1016/j.future.2019.02.011

  Main contributions: design, development and integration of the AAA module of the BIGSEA platform.

- Alic A. et al. (2018). GIS and Data: Three Applications to Enhance Mobility. In XIX Brazilian Symposium on Geoinformatics - GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, (pp. 1–12).

  Main contributions: design, development and integration of the AAA module of the BIGSEA platform.

Seminars

- Silva P. (2019). Privacy Enhancing Technologies - Protection and Control of Secured Information in the Cloud. Ostrava University, Ostrava, Czech Republic, September 17.
- Silva P., Casaleiro R. (2019). Risk Management and Data Analysis in PoSeID-on. Workshop on Privacy, Data Protection and Digital Identity, University of Coimbra, Coimbra, Portugal, July 11.
- Silva P. (2019). Privacy Enhancing Technologies - Protection and Control of Secured Information in the Cloud. Ostrava University, Ostrava, Czech Republic, April 4.
- Silva P. (2017). Users' Education Regarding Safer Online Behaviors and Data Privacy Protection. EB23 Dr.ª Maria Alice Gouveia, Coimbra, Portugal, March 30.

Contents

List of Figures

List of Tables

Acronyms

AAAaaS Authentication, Authorisation and Accounting as a Service

AAA Authentication, Authorisation and Accounting

AD Active Directory

AI Artificial Intelligence

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

CA Certificate Authority

CPU Central Processing Unit

CISUC Centre for Informatics and Systems of the University of Coimbra

CPPDPA Computer Processed Personal Data Protection Act

CNNs Convolutional Neural Networks

CRFs Conditional Random Fields

DL Deep Learning

DIR Department of Information Resources

DNA Deoxyribonucleic Acid

DoD Department of Defense

DOJ Department of Justice

DP Data Processor

DUI Driving Under the Influence

EC European Commission

ECC Elliptic Curve Cryptography

ELMo Embeddings from Language Models

ETL Extract Transform Load

EU European Union

FL Federated Learning

FP False Positives

FN False Negatives

GDPR General Data Protection Regulation

GLBA Gramm-Leach-Bliley Act

GUI Graphical User Interface

GMB Groningen Meaning Bank

HIPAA Health Insurance Portability and Accountability Act

HTML Hypertext Markup Language

HMM Hidden Markov Models

IAM Identity and Access Management

IDaaS Identity as a Service

IdP Identity Provider

ISO International Organisation for Standardisation

IBAN International Bank Account Number

JSON JavaScript Object Notation

LCT Laboratory of Communication and Telematics

MLP Multi-Layer Perceptron

ML Machine Learning

NER Named Entity Recognition

NLP Natural Language Processing

NLU Natural Language Understanding

NLTK Natural Language Toolkit

NoSQL Not only Structured Query Language

OTP One-Time Passwords

PDA Personal Data Analyser

PDF Portable Document Format

PDPA Personal Data Protection Act

PET Privacy Enhancing Technology

PETs Privacy Enhancing Technologies

PII Personally Identifiable Information

PIPA Personal Information Protection Act

PIPEDA Personal Information Protection and Electronic Documents Act

POS Part of Speech

QoS Quality of Service

REST Representational State Transfer

RF Random Forest

RoBERTa Robustly Optimized BERT Pretraining Approach

RSA Rivest–Shamir–Adleman

RMM Risk Management Module

SAML Security Assertion Markup Language

SSL Secure Sockets Layer

SSO Single Sign-On

SSD Solid State Drive

TP True Positives

TN True Negatives

ULMFiT Universal Language Model Fine-tuning for Text Classification

US United States

USA United States of America

XML Extensible Markup Language

Chapter 1

Introduction

Contents

This thesis addresses the importance of safeguarding the access and safekeeping of individuals' private information. A combination of security and privacy mechanisms are proposed to accomplish this goal and minimise the identified issues. The process included the definition and development of an Authentication, Authorisation and Accounting as a Service (AAAaaS) module, a hybrid approach that recurs to Natural Language Processing (NLP) tools and Machine Learning (ML) to monitor and classify Personally Identifiable Information (PII), as well as mechanisms that improve individuals' privacy awareness and minimise privacy risks.

The research background and motivation of this thesis are presented next. They are followed by the objectives and respective contributions, as well as the thesis outline.

1.1 Motivation and Problem Statement

Over time different technologies and solutions have been proposed to secure users' information online and offline. These solutions range from privacy policies to security mechanisms, including encryption, authentication methods, anonymisation techniques, laws, and regulations. All these solutions play an essential role in providing proper data privacy protection and security to users' information in the Cloud.

Traditional authentication systems (e.g., password-based authentication) are among the most common and widely used methods of securing access to data, systems, databases, or services [Bellovin et al., 2003]. Nevertheless, authentication systems can be subject to attacks or can fail. An example is the JPMorgan attack [Kirk, 2014], which resulted in the exposure of personal information (e.g., names, addresses, email) that compromised 87 million customers. Uber was also a target of an attack [Dunn, 2018] and information about 57 million customers, as well as drivers, was compromised. Although virtually any system may be subject to previously unknown exploits or failures, the likelihood of these events could be minimised by adopting suitable security and privacy mechanisms as the ones proposed in this thesis. Further details about privacy threats are discussed in Chapter 2.

Recent data breaches and privacy scandals (shown in the next Chapter) have also triggered discussion, more specific policy-making and further research within the privacy area. In turn, they have lead to national and regional legislation, such as European Union (EU)'s General Data Protection Regulation (GDPR) [Schulz and Hennis-Plasschaert, 2016], that aim at providing legal assurances in what concerns the protection of PII.

Cloud Computing and the associated services and applications are every day more involved in our digital lives. The implications are significant, as massive amounts of data are being generated and held online every day [Marr, 2018]. Therefore, data privacy should be a requirement and fundamental characteristic of offline processing and online services in the Cloud.

In Europe, civil society, academia, industry, and policymakers are driven by GDPR-compliance [Schulz and Hennis-Plasschaert, 2016], as well as its practical and legal effects. In other regions, other regulations are applied. For instance, Gramm-Leach-Bliley Act (GLBA) or Health Insurance Portability and Accountability Act (HIPAA) in the United States of America (USA) [Federal Deposit Insurance Corporation, 2019; US Code, 1999; Mercuri, 2004]; Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada [Singh and Chatterjee, 2017]; Personal Data Protection Act (PDPA) in Russia [Sergey, 2018]; or Computer Processed Personal Data Protection Act (CPPDPA) and Personal Information Protection Act (PIPA) in China [Greenleaf and Chen, 2012]. Additional details about privacy regulations can be found in Chapter 2.

The widespread use of digital services has led to individuals' concerns on security and privacy [Yun et al., 2019], as well as on the processing of their personal information by data processors and third parties. On the other hand, technological advancements continue to deliver services, tools, and applications that are increasingly demanding of PII. These demands are justified for the sake of data analytics, to drive businesses, and generally to enhance user experience [Yoo, 2010]. This is applicable in a large variety of areas: public administrations, health care, business and many others. In this context, demand for novel and effective ways of protecting and controlling PII has never been so high [Domingo-Ferrer et al., 2019].

It is known that security concerns of Cloud systems or services should not be addressed individually or in an ad-hoc manner, as this may result in insufficient solutions [Khalil et al., 2013]. Security should follow appropriate standards, be dully tested and capable of being integrated into state-of-the-art environments. Moreover, Cloud systems require mechanisms that ensure security and privacy, as well as scalability and efficiency [Tari et al., 2015].

In light of the issues mentioned above, security enhancements can be achieved with the proposal of a pluggable cloud-based Authentication, Authorisation and Accounting (AAA) mechanism. An AAAaaS, proposed in this thesis, is cloud-oriented and provides the general functionalities of traditional AAA and Identity and Access Management (IAM) services. Additionally, it includes interfacing with external identity providers using OAuth [Hardt et al., 2012]. The solution is deployable and manageable according to three fundamental Cloud principles: scalability, elasticity and resilience. Additional details can be found in Chapter 3.

Artificial Intelligence (AI) mechanisms such as Named Entity Recognition (NER) and other ML algorithms can play an important role on the privacy front. Instead of performing data analysis for the benefit of businesses and organisations, one can consider approaches that directly benefit the user. The mechanisms proposed in this thesis show that performing transparent data analysis for the sole benefit of the user, is an effective and privacy-preserving way of monitoring PII shared with or between third parties. Further details are provided in Chapter 4.

Privacy risks depend on a variety of factors and scenarios [De Joyee and Le Métayer, 2016]. Nevertheless, some aspects should be considered in most cases: data types involved, sensitivity, correlation, user consent, retention time and data processor reputation. Therefore, this thesis proposes a multi-input privacy risk assessment mechanism that considers those six aspects. As the mechanism is integrated with a real-world scenarios (i.e., various pilots of the PoSeID-on project), every time privacy thresholds are triggered, data subjects receive privacy warnings. A complete specification of this mechanism is provided in Chapter 5.

1.2 Objectives and Contributions

The primary goal of this thesis is to enhance security and privacy in Cloud environments. Such objective was possible by proposing a cloud-based AAA mechanism, an ML automated data analysis pipeline and a multi-input privacy risk assessment mechanism. This is accompanied by a literature review of the state-of-the-art of ML mechanisms, with a focus on NER and PII identification.

The proposed mechanisms are also integrated, evaluated, and validated in two different international projects: EUBra-BIGSEA and PoSeID-on, providing results closer to real-world implementations.

The specific objectives of this thesis are as follows:

Objective 1 - Provide a method for authorisation, authentication and accounting in Cloud environments.

Objective 2 - Propose automatic and AI-based mechanisms for personal data analysis and classification in compliance with existent regulations for the sole benefit of the user.

Objective 3 - Design and build a multi-input model for privacy risk assessment in Cloud environments.

Objective 4 - Develop and integrate the proposed mechanisms in Cloud platforms to evaluate them under real-world scenarios.

Taking into consideration the specific goals, this thesis has produced the following main contributions:

Contribution 1, Design and development of an AAA mechanism

This contribution relates to the design and development of an elastic and efficient AAA security module able to cope with the requirements of Cloud systems. Easy to configure and to deploy, this contribution provides support to both infrastructure and applications. The possibility of interfacing with external identity providers also enhances the potential of this service. This contribution is presented in Chapter 3.

Contribution 2, Named Entity Recognition models for PII classification

A method based on ML algorithms such as Multilayer Perceptron (MLP) and Random Forests (RF) that detects privacy violations through the identification of personal information. The process involves pre-training word vectors, using publicly available data sources, labelling entities, as well as training and evaluating models. The resultant Named Entity Recognition (NER) models achieved high $F_1 score$ in the classification of the entities. This contribution is described in Section 4.4. Prior to that, a comparative study of the features and performance of existent NLP tools was essential. The details can be found in Section 4.3.

Contribution 3, Hybrid data classification mechanism

This contribution proposes a hybrid NER pipeline for detection of personally identifiable information. This mechanism includes the results from Contribution 2 as well as NLP tools and regular expressions. The outcome is an ensemble mechanism that collects the output from all the classifiers and infers which entity best matches the data. The main objective of the mechanism is not to provide a perfect solution, which is hard in NLP, but rather a practical way of dealing with the problem and allowing the fine-tuning of the weight of the mechanisms involved. This contribution is presented in Section 4.5.

Contribution 4, Multi-input privacy risk assessment mechanism

Conceptual definition and implementation of a multi-input mechanism to assess privacy risks in data transactions. The inputs include data sensitiveness, correlation, retention time, data validation, and reputation of the involved parties. A fuzzy logic classifier and respective fuzzy rules and sets are also applied in the process. The mechanism provides a normalised quantification of the privacy risk level associated with data transactions. This contribution is evidenced in Chapter 5.

Contribution 5, Automatic mechanism for privacy-enhancing data analysis

The proposal, design and development of a privacy-enhancing mechanism capable of automatic data analysis. This contribution benefits from data classification mechanisms considered in Contributions 2 and 3, as well as the privacy risk assessment mechanism mentioned in Contribution 4. The proposed mechanism leverages a combination of data analysis and privacy risk assessment mechanisms to compute privacy risk levels for the sole benefit of data subjects. Then, it issues different levels of privacy warnings, raising user awareness and minimising privacy threats. This contribution, built upon the contributions evidenced in Chapter 4 and 5, results in the Personal Data Analyser (PDA) module.

Contribution 6, Impact assessment of the proposed mechanisms in real-world Cloud platforms

All methods and mechanisms proposed in this thesis (available in container images) are deployed, integrated and validated in Cloud platforms. Those platforms are based on Mesos [Apache Mesos, 2020], Marathon [Mesosphere, Inc., 2020], and Kubernetes [The Linux Foundation, 2020] orchestration mechanisms. Both AAAaaS (Contribution 1) and PDA (Contribution 5) were validated by users of applications and services deployed in the EUBra BIG-SEA and PoSeID-on Cloud platforms. A positive feedback was collect across the board, which demonstrates the real impact the proposed mechanisms have beyond the theoretical demonstration.

1.3 Outline of the Thesis

This thesis is organised as follows.

Chapter 2 - Privacy Enhancing Technologies

Introduces different privacy concepts, technologies, threats and regulations, and discusses specific data protection and privacy challenges in the Cloud.

Chapter 3 - Authentication, Authorisation and Accounting in the Cloud

Addresses essential security requirements of any system: authentication, authorisation and accounting. It describes the design, development and integration steps taken towards a cloud-based AAA mechanism.

Chapter 4 - Advances on Automated Personal Data Analysis

Discusses how Machine Learning mechanisms are used for personal data monitoring and classification for the solo benefit of the user. In this chapter, aspects such as Neural Networks, Random Forests, Natural Language Processing, Named Entity Recognition, and personal data analysis are explored as part of the proposed privacy-preserving data analysis mechanisms.

Chapter 5 - Privacy Risk Assessment Mechanisms

Proposes a fuzzy logic-enabled multi-input privacy risk model capable of processing discrete or categorical elements such as data sensitiveness, correlation, retention times or data processor reputation. Membership functions assess how the different elements influence privacy risks.

Chapter 6 - Conclusions and Future Work

Presents the final remarks and conclusions, as well as the outline of future research to further advance this work.

Chapter 2

Privacy Enhancing Technologies

Contents

Privacy Enhancing Technologies (PETs) are mechanisms and technologies designed to protect personal data in different ways and stages of its life cycle (e.g., in transit, at rest, in or off-premises).

This chapter provides a background on Privacy Enhancing Technologies, introduces the different privacy concepts, technologies, threats and regulations, and discusses specific data protection and privacy challenges in the Cloud.

The literature revision presented in this chapter intends to provide the reader with the necessary background about the topics discussed in this thesis. A more extended background and analysis of PETs is provided in the following paper:

- Silva, P. et al. (2020). Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges. IEEE Access, Vol. 9, pp. 10473-10497, Print ISSN: 2169-3536. Online ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3049599, January 2021.

2.1 Introduction

Before the increase of Cloud Services, the Internet already had an abundance of services that required data protection mechanisms. As Cloud Computing related services emerged and spread, privacy concerns were raised as more sensitive data was outsourced to the Cloud. That was due to the Cloud's intrinsic characteristics, such as distributed online storage, data replication, data integration, data regulation in different countries, privacy policies, and different types of threats.

On the one hand, robust security features like state-of-the-art cryptography or privacy-enhancing data analytics with Machine Learning are essential to ensure proper data access, management and protection. On the other hand, sophisticated data masking techniques are pivotal to transform data in such a way that allows value extraction in scientific or commercial contexts without disclosing sensitive information. The mechanisms proposed in this thesis are designed to leverage the capabilities of AI mechanisms while respecting privacy regulations and avoiding privacy threats.

The next section presents the different concepts and applicability domain of Privacy Enhancing Technologies.

2.2 Concepts and Applicability Domains

Nowadays, the word *privacy* can be ambiguous and therefore difficult to accurately define. There are several forms and definitions of privacy, none of them less relevant. In simple terms, to have privacy is to have the ability to control which personal information is known, used and how it is accessed. Personal information is every piece of information that is related to an identifiable person

[Barth et al., 2006]. The following concepts represent common expressions and keywords used in the field:

- *Anonymization* – Daintith defines anonymization as "a process that removes or replaces identity information from a communication or record" [Daintith, 2019]. For instance, a subject in communications or records can be made pseudonymous. The same subject will then always have the same replacement identity but cannot be identified.
- *Authentication* – According to [Peisert et al., 2013], authentication is the process of confirming the identity of a user and can be seen as the degree of trust that one can have that the source of data is who it claims to be. This also applies to machines or services.
- *Authorisation* – It defines the extent of access to a system and what type of services and resources are accessible by the authenticated entity.
- *Accounting* – It is the act of registering what was done in the system (e.g., login or logout).
- *Concealing* – Petitcolas et al. [Petitcolas et al., 1999] state that concealing is the act of keeping from sight, to hide. By doing so, it means to keep something secret or prevent something from being known or noticed.
- *Data Confidentiality* – According to the Oxford dictionary [Wiles et al., 2008], something confidential is: "intended to be kept secret," meaning that confidential information is the information intended to be kept secret. It can be seen as a set of rules that limit access or impose restrictions on certain types of information. Thereby, providing data confidentiality means keeping data secret.
- *Data Curator* – A data curator is an individual in charge of managing data. As Cragin et al. state: "Data curation is the active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time" [Cragin et al., 2007].
- *Data Privacy* – Data privacy is the ability of an individual or group to stop information about themselves from becoming known to people, other than those whom they choose to give the information to. Privacy is sometimes related to anonymity, and Solove [Solove, 2009] considers that it is often most highly valued by people who have private data publicly known.
- *Data Utility (or Data Usability)* – After the anonymization process, there is the matter of the utility of the information, which is of high importance. Sweeney [Sweeney, 2002] considers utility or usability as the representational value of the amount of information preserved in the anonymized data.
- *De-identification* – De-identification is the process of identification, selection, and removal of sensitive information in a document or data set

[Tomashchuk et al., 2019].

- *Observable Data* – The information that is available for a limited amount of time. In this case, an attacker might need to be present to observe or collect the data. Examples are communication systems where contents or intervening parties are actively or passively compromised.
- *Personally Identifiable Information* – Krishnamurthy and Wills [Krishnamurthy and Wills, 2009] define PII as the information which can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linkable to a specific individual.
- *Published Data* – Published data is all the information willingly released and available to the public, considering all formats: databases, logs, traces, social network profiles, posts, and others.

It is also important to fully understand the kind of data to which these concepts can be applied. Many areas hold Personally Identifiable Information (PII) by default. Those areas include, but are not limited to, *health care*, *criminal*, *financial*, and *social* information. *Health care* information is one of the most sensitive types as it relates to an individual's health record. Blood samples, urine, Deoxyribonucleic Acid (DNA), and saliva test results are examples of health information as they relate to biological and genetic profiles, regardless of the origin. *Criminal*-related information can range from criminal records to court rulings, charges, convictions, speed tickets, Driving Under the Influence (DUI) of alcohol or drugs, and many other associated records. *Financial* information regards all information related to an individual's finances, such as salary, debt, mortgage, and other records such as bank accounts, credit and debit cards, bank extracts, loans, leases, and taxes. *Social* information includes, for instance, name, address, marital status, family, gender, sexual orientation, education, voter information, political preferences, location data, shopping habits and many others.

The Cloud comprises an enormous amount of information stored or transmitted online. It can be processed in various locations, at times with unclear information about the duration of data collection and often with no guarantees of permanent deletion options. Along with all sorts of personal information or media like image and video stored in social networks or applications and web services, there are online communication services such as email. An example is a company processing email contents to provide targeted advertising or personal assistant-related features. There are other aspects, such as shopping habits, product preferences, interaction and communication with others, and many others. What usually applies in most cases is that most online users leave a track, thus forming a digital fingerprint that can lead to complete or partial identification. Location, browser, search queries, visited websites, cookies, canvas, and window size are examples of data used to identify users.

Architecture and Design, *Communications and Networking*, *Data Management* and, *Identify Management*, depicted in Figure 2.1, are four applicability domains where Privacy Enhancing Technologies (PETs) can be of use. This is justified

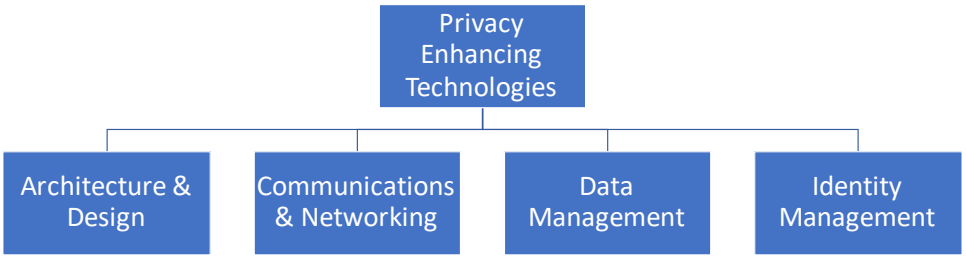


Figure 2.1: Proposed applicability domains of PETs.

by the coverage that these four aspects provide: (1) the architecture and design of applications with privacy embedded by design; (2) providing secure networks and private communications; (3) keeping private the information available in the most variate data types; (4) keeping users' identity private.

Therefore, following a development approach that applies sate-of-the-art mechanisms and methodologies in these four application domains should result in a privacy-assuring product or service. Two of the domains, Data and Identity Management, are the focus of this thesis. Nevertheless, the mechanisms proposed in this thesis are compliant with the other domains as they were validated under platforms and environments that fall under the first two domains.

The next section presents different types of privacy threats and attack models.

2.3 Privacy Threats

An invasion of privacy occurs when personal information is used without consent or knowledge of the owner. It can happen through a data breach, attack, eavesdropping, or other forms of appropriation. According to Drake [Drake, 2019], Robison [Robison, 2018] and Thomson Reuters' Find Law [Schonrock et al., 2018], privacy threats (as shown in Figure 2.2) can be classified as follows:

- *Intrusion* – An intrusion of privacy includes all the actions that directly or indirectly invade an individual or organisation's private affairs. Phone calls or conversations recorded without authorisation and knowledge, taking pictures or trespassing on private property, repeatedly making non-requested phone calls, or spying on someone are examples of privacy intrusion.
- *Public Disclosure* – Releasing previously unknown or private information to the public is a public disclosure. This information can be offensive or embarrassing when publicly released. Therefore, if the data does not provide any public concern, the one(s) responsible for the release can be liable for privacy invasion. Typical examples are individuals in public of-

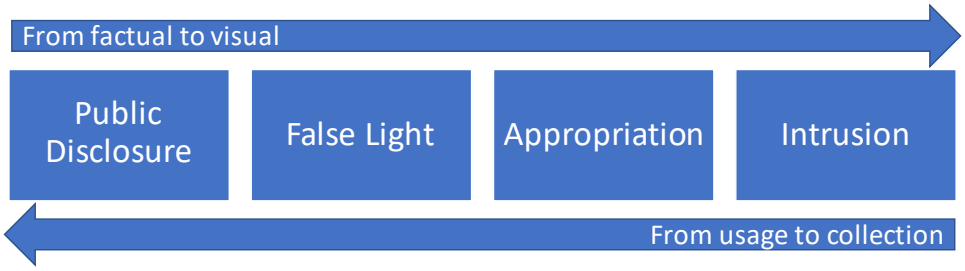


Figure 2.2: Privacy threats categories.

fices, celebrities, or politicians that have their private information publicly disclosed.

- *False Light* – Similar to the previous point (public disclosure) is false light. It is a form of public disclosure of false or malicious statements. It is usually done by distorting the truth or using fictional facts.
- *Appropriation* – This case refers to the appropriation of an individual or organisation's name or identity. It usually happens by using an individual's name, image, or any other personal characteristic without authorisation or knowledge. It is common to see such media cases, references in books, stories, or marketing. Although it is possible to happen with any person, the issue is more recurrent with celebrities or famous personalities. In the digital era, this happens with online profiles or accounts as well.

There are additional types of privacy threats. For instance, Solove [Solove, 2009] proposes, a similar, yet more fine-grained taxonomy: information collection (e.g., surveillance), information processing (e.g., identification or re-identification), information dissemination (e.g., disclosure) and invasion (e.g., decisional interference). Other types of privacy invasion are attacks directed to data records. As defined by the International Organisation for Standardisation (ISO), an attack is an "attempt to destroy, expose, alter, disable, steal or gain unauthorised access to or make unauthorised use of anything that has value" [ISO, 2009] to an individual or organisation.

Within the data privacy scope, the overall consensus is that there are three different ways an attacker gathers information (i.e., attacker estimates) [Gachanga et al., 2018; Zaman et al., 2017; Park et al., 2018]. These attacker estimates are based on the type of information available to an attacker and the resemblance with other gathering information methods. The three main attacker estimates are:

- *Prosecutor* – The attacker knows that data about the targeted individual is contained in the data set.
- *Journalist* – The attacker has no background knowledge.
- *Marketer* – The attacker is not interested in re-identifying just a specific individual.

It is also possible to enforce particular attack models that operate on specific

data conditions. The attack models identified by Fung et al. [Fung et al., 2010] are the following:

- *Record Linkage* – It happens when an attacker successfully matches a record owner to a sensitive attribute from datasets published or obtained elsewhere.
- *Attribute Linkage* – When there is no specific record identification but the attacker can still infer sensitive values supported by the information of the group where the record owner belongs.
- *Table Linkage* – When attacks successfully derive the presence or the absence of the targeted record owner in a table.
- *Probabilistic Attack* – Based on the uninformative principle from Machanavajjhala et al. [Machanavajjhala et al., 2007]. Instead of focusing on actual records, it assures that the beliefs before and after accessing published data do not change significantly.

Table 2.1 provides examples of privacy invasions in which private data was exposed, and citizens' or organisation's privacy was compromised. Political interests, credit card details or addresses were publicly disclosed. In some cases, like Yahoo and Uber, data breaches happened due to security reasons. However, in other cases (such as Netflix or AOL), it was due to the incorrect usage of anonymization mechanisms. Cross-referencing or linkage attacks pose a significant risk for anonymized data. Nevertheless, the risk can be minimised or possibly avoided if proper anonymization mechanisms - ideally, a combination of mechanisms - are used, and attacker models are considered.

The next section presents several privacy regulations available in different regions. Moreover, it discusses their applicability to different data types.

Table 2.1: Examples of previous privacy breaches and exploited threats.

Who?	**What?**	**How?**	**When?**	**Threat Exploited**	**Source**
AOL	Published search data led to identification of users	Cross referencing	2006	Public Disclosure	[Butler, 2007]
Netflix	Released data sets led to identification of users	Cross referencing	2006	Public Disclosure	[Narayanan and Shmatikov, 2006]
Yahoo	500 million user accounts stolen	Hacking	2014	Intrusion and Appropriation	[Trautman and Ormerod, 2016]
JPMorgan	87 million customers details exposed	Hacking	2014	Intrusion and Appropriation	[Kirk, 2014]
Uber	57 million customers and drivers details exposed	Hacking	2016	Intrusion and Appropriation	[Dunn, 2018]
Equifax	Sensitive information of 140 million people	Hacking	2017	Intrusion and Appropriation	[Burns and Johnson, 2018]
Cambridge Analytica	Unauthorised profiling	Personal data scraped from Facebook accounts	2018	Intrusion	[Isaak and Hanna, 2018]
Facebook	Over 540 million records exposed	Third-party security issues (Cultura Colectiva)	2019	Intrusion and Appropriation	[UpGuard, 2019]
Microsoft	Over 250 million records exposed	Security issues	2020	Intrusion and Appropriation	[Cimpanu, 2020]

2.4 Privacy Regulations

Many countries have laws and regulations regarding privacy, data access, data sharing, or handling. In Europe, some directives should be enforced and/or followed by the countries that are part of the EU. Moreover, with the GDPR enforcement since May 2018, any services or businesses handling data from European citizens are forced to comply with this regulation. In the USA the GLBA [Federal Deposit Insurance Corporation, 2019; US Code, 1999] is being enforced, while in Canada there is the PIPEDA [Singh and Chatterjee, 2017]. To the East, there is the Russian Federation with its PDPA [Sergey, 2018]. Regarding China, Graham and Hui-ling [Greenleaf and Chen, 2012] show that although there is no national privacy law enforced, the CPPDPA and PIPA are examples of regulations created for that effect.

Although regulations vary from country to country, they have a common objective: to provide legal protection and regulation over its citizens' personal and private information. The particularities of the regulations in the USA and Europe are analysed next.

In the USA different activity sectors (e.g., insurance, financial, or health care) have their own regulations. The three main regulations are the following:

- *HIPPA* – The Health Insurance Portability and Accountability Act is a health care regulation that assures that individuals' health information is properly protected while still: (1) simplifying administrative processes by standardising health care transactions; (2) reforming insurance conditions so that a job change does not affect coverage. Failure to comply with this regulation can result in fines up to $ 250K [Mercuri, 2004] and up to 10 years of jail time.
- *GLBA* – The Gramm-Leach-Bliley Act regulates how financial institutions manage financial information. Banks, insurance companies, securities firms, or even retailers, must provide confidentiality about customers' credit information. Furthermore, according to the Federal Deposit Insurance Corporation [Federal Deposit Insurance Corporation, 2019] and the U.S. Code [US Code, 1999], these institutions must inform their customers how their information is kept confidential and secure.
- *CLOUD* – The Clarifying Lawful Overseas Use of Data Act regulates authorities' access to data held by American companies across the border of the USA. The act allows the Department of Justice (DOJ) data access without authorization from the courts or the Senate [Senate of the United States, 2018; Moon, 2018].

While in the USA there is a sectoral approach for privacy regulation, in the EU the GDPR regulates citizens' data privacy transversally with regard to all types of personal information [Schulz and Hennis-Plasschaert, 2016]. Some of the key points of the GDPR are as follows:

- *Territorial Applicability* – This point is directed to all companies that process the personal data of European Union residents, regardless of the

company's location.

- *Penalties* – The applicable penalties are up to 4% of annual sales volume or a maximum of € 20M. This penalty is applied in severe cases (for instance, lacking customer consent to process data).
- *Consent* – All consent requests must also be given in an easily accessible form. The purpose of data processing should also be present in the consent request.
- *Right to Access* – This right intends to provide citizens with access to copies of all personal data held by a company. Furthermore, it is the right to know whether their data is being processed, the purpose, and the location.
- *Breach Notification* – It is mandatory to issue a breach notification (with a 72-hour limit) in cases where the data breach can pose a risk for the rights and freedom of citizens.
- *Right to be Forgotten* – The right to be forgotten gives the right of having a citizen's data erased. It also has the potential to prevent data processing from third parties.
- *Data Portability* – This option grants a citizen the right to receive and transmit his / her data.
- *Privacy by Design* – PbD is the inclusion of privacy and data protection mechanisms at each stage of development of a system or service, rather than addition. Companies such as Microsoft already adopt this principle when developing new products or services [Microsoft, 2014].
- *Data Protection Officers* – DPOs are mandatory for those whose core activities consist of processing operations that require regular and systematic monitoring of data subjects on a large scale, particular categories of data, or data relating to criminal convictions and offences.

Since 2000 there had been an agreement concerning privacy between the European Commission (EC) and the USA Government: the Safe Harbor agreement [U.S. DoC, 2000]. The primary purpose was to prevent and avoid accidental disclosures of personal information.

Despite the enforcement of such an agreement, after an EU citizen complained about Facebook's handling of his data, the agreement was declared invalid [Gibbs, 2015] by the European Court of Justice. After a modification of data collection terms between the USA and the EU, a new agreement was drafted: the EU-USA Privacy Shield. It is described as a framework for transatlantic exchanges of personal data for commercial purposes between the EU and the USA, and it is designed to accommodate the European regulations.

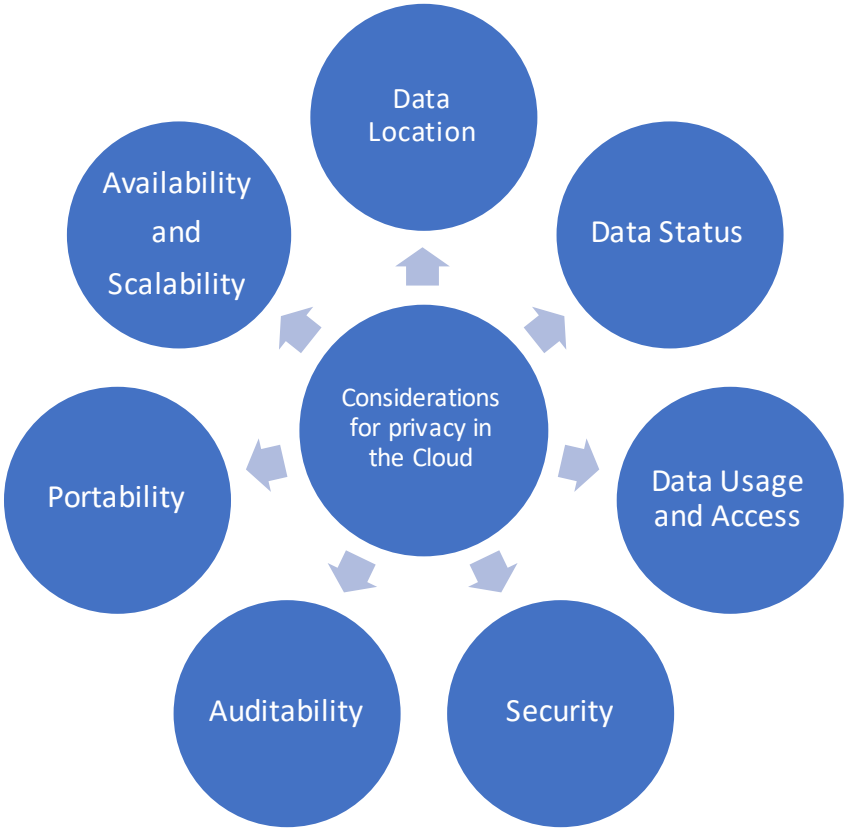


Figure 2.3: Considerations for privacy in the Cloud.

2.5 Privacy in the Cloud

Cloud Services differ from more traditional Internet Services. The distributed data processing or the servers' location are aspects to consider in regard to privacy. All the aspects discussed in the previous sections should be suitable and adapted to the Cloud's context. The main requirements for privacy in the Cloud are the following:

- *Data Location* – Privacy laws and regulations differ from country and region. Therefore, compliance in different locations is a challenge. Companies processing data from international customers (e.g., European or American citizens) face some difficulties since the servers with databases and computing power might be distributed across different countries. There are at least two aspects to consider: local laws regarding the storage and management of customers data (e.g., GLBA) and laws regarding the country of origin of the customer's data (e.g., GDPR). Failure to comply might incur in significant losses (i.e., fines) for the companies in question.
- *Data Status* – Another aspect that Cloud service providers should consider is the disclosure of the methods used to protect data (e.g., the disclosed privacy policy). The status of the data during the stage(s) of processing or handling should be indicated (e.g., plain text, encrypted, anonymized, or pseudonymized).

- *Data Usage and Access* – It is necessary to assure proper handling and access to data at all times. A system or service might be compromised even if appropriate security measures and policies are in place (e.g., hacks on the U.S. Treasury and Commerce departments [Bing, 2020], and SolarWinds [Lambert, 2020]). Ensuring proper data usage policies and (both physical and logical) access is sometimes not given due diligence. Suppose a more specific data processing is intended. In that case, it is recommended to disclose usage policies in two directions: user/customer to service provider, and service provider to user/customer (nevertheless, the latter predominates in most cases).

  Regarding data access, it is crucial to accurately define access rules. As such, a series of questions can be addressed: who can access, why, how, where, and for how long? These questions can be answered with AAA mechanisms such as the one described in Chapter 3.

- *Security* – In addition to data status, there is a security point of view. In this case, infrastructure (e.g., an Intrusion Detection System [Liao et al., 2013]), communications (e.g., a Secure Sockets Layer [Hickman and Elgamal, 1995]), and other security features play a crucial role in keeping data secure, regardless of the data state (e.g., plain text or encrypted). Common aspects such as establishing strong passwords, antivirus, and regular software updates can effectively increase security on both ends: users/customers and service providers.

As depicted in Figure 2.3, other requirements such as auditability, portability, and availability, should also be considered. Nevertheless, there might still be vulnerabilities despite the Cloud providers' active measures to meet high privacy standards. Typically, data owners or users of such services have no physical control over the system. Therefore, instead of full-trust, there is a semi-trust relationship. Nevertheless, in cases where Cloud Services are used for the single purpose of outsourcing data (i.e., data storage), users may take more proactive approaches such as anonymizing their data. For that purpose, several privacy algorithms and tools can limit the exposure of sensitive information.

2.6 Summary

This chapter started by providing a background on the main concepts and applicability domains of Privacy Enhancing Technologies (Section 2.2). It also identified the privacy threats most commonly observed in Cloud environments (Section 2.3). A summary of existent regulations (Section 2.4) and an analysis of privacy mechanisms for the Cloud (Section 2.5) are equally provided.

The literature revision presented in this chapter intended to provide the reader with the necessary background about the topics discussed in this thesis. An extended background and analysis of PETs is provided the following paper:

- Silva, P. et al. (2020). Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges. IEEE Access, Vol. 9, pp. 10473-

10497, Print ISSN: 2169-3536. Online ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3049599, January 2021.

The following chapters present mechanisms that contribute to personal data protection and increased privacy preservation in Cloud environments.

Chapter 3

Authentication, Authorisation and Accounting in the Cloud

Contents

An essential security requirement of any system is the correct identification and access management of users, making it crucial to have means to univocally authenticate and authorise users. This can be accomplished by providing adequate Authentication, Authorisation and Accounting (AAA) mechanisms.

This chapter describes the design and development of a cloud-based AAA mechanism provided as-a-service (AAAaaS). This mechanism was integrated and validated in the scope of the H2020 EUBra-BIGSEA Project.

The outcomes of this chapter resulted in the following publications:

- Alic A., et al. (2019). **BIGSEA: A Big Data Analytics Platform for Public Transportation Information**. Future Generation Computer Systems, 96, 243 - 269. DOI: 10.1016/j.future.2019.02.011
- Silva P. et al. (2018). **A Europe-Brazil Context for Secure Data Analytics in the Cloud** in IEEE Security & Privacy, vol. 16, no. 6, pp. 52-60. DOI: 10.1109/MSEC.2018.2875326.
- Alic A. et al. (2018). **GIS and Data: Three Applications to Enhance Mobility**. In XIX Brazilian Symposium on Geoinformatics - GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, (pp. 1–12).

3.1 Introduction

Cloud services are naturally designed to be remotely accessed. Therefore, it is necessary to have means to univocally authenticate users. Cryptography should be present in all the steps of this process. For instance, using certificates or digital signatures is a simple yet effective way of protecting sensitive information in transit or at rest. After user authentication, the system must still maintain control over which operations the user can perform. In other words, it is necessary to provide authorisation mechanisms, so that the system can verify which services are allowed for a specific user. These operations can be achieved by implementing AAA mechanisms [Perkins and Calhoun, 2005].

The first and second *A*'s are more commonly used, but the last, which stands for "accounting", is usually ignored during the design of security solutions. However, it is also a relevant requirement. It is responsible not only for collecting information that allows further investigation in the case of attacks, but also for providing precise knowledge about resources utilisation.

3.1.1 Background

AAA services address the need for controlling and managing access to data, services or system resources, which can be synthesised in three key questions. *Authentication* wants to know who or what is trying to access the system, while *authorisation* focuses on knowing what you can do in this system. Finally, *accounting* registers what is done in the system.

Due to the criticality of the data stored, managed, and analysed in most systems, AAA services/mechanisms have to provide appropriate access control. Also, the link between the external services/users and application development services is a critical aspect, as it allows for external access points to the platform. These services are transversal to all components and should enforce additional security mechanisms.

Different security methods can complement the mechanisms mentioned above. Cryptography, for instance, allows two persons or services to securely communicate over an insecure channel subject to eavesdroppers [Coron, 2006]. Cryptography is the science that studies techniques and methods for secure communications by using secrets or hide information. Rivest–Shamir–Adleman (RSA) [Rivest et al., 1978] is one of the first cryptography algorithms using a public key. It was designed based on the difficulty of factoring the product of two large prime numbers. This algorithm complements other methods that offer safe and private methods of authentication, such as digital certificates. Elliptic Curve Cryptography (ECC) [Hankerson et al., 2006] is another cryptography solution that operates under elliptic curves and finite fields. ECC provide shorter keys than RSA, which in some cases is the reason why organisations opt for ECC instead of RSA.

A digital certificate (e.g., X.509 [Housley et al., 2002]) is a record that holds information about a user or service. This information is presented in a way that its authenticity and truth can be verified by a Certificate Authority (CA). To ensure that data is verified by a CA, it must use cryptography methods. The CA uses digital signatures to provide a robust method that allows verification of the information in the certificate. When a certificate is issued by a CA and exchanged by the two parties involved, all communications are encrypted using Secure Sockets Layer (SSL) [Hickman and Elgamal, 1995]. By creating a secure communication channel between two parties, SSL encrypts all the information exchanged between them. This a way of assuring the integrity of messages, since all received content must be exactly like what was sent. Biometrics such as fingerprints, face scans, hand geometries and other types of personal identification, can also be used to verify the parties communicating or involved in a transaction. However, there are long-lasting concerns [Woodward, 1997; Schneier, 1999] about the risks of using biometrics in such contexts as it is usually linked with storing that information and keeping individual's most sensitive information.

For organisations seeking a way to authenticate users to their on-premises and cloud-based resources, an Identity Provider (IdP) or a AAAaaS service (as proposed in this Chapter) are suitable solutions. It is possible to point out several tools and mechanisms that can be used to offer these features. In particular, there are plenty of authentication mechanisms not only proposed in literature but also extensively implemented with different technologies and platforms. The next section describes some of those mechanisms.

3.1.2 Related Work and Open Issues

Although conventional AAA solutions are available, like for example Radius [Rigney et al., 2000], Diameter [Fajardo et al., 2012] and TACACS+ [Carrel and Grant, 1997], not many complete solutions are constructed in the context of Cloud computing. One such solution is OpenID [Recordon and Reed, 2006], which offers identification mechanisms and has broad usage, making it an interesting option.

There are mechanisms and standards such as Security Assertion Markup Language (SAML) [Cantor et al., 2004] and OAuth [Hardt et al., 2012] that allow the exchange of authentication and authorisation information data between entities. This information exchange is also known as Single Sign-On (SSO) [Armando et al., 2008], which allows delegating the authentication functionality to external identity provider services like the ones offered by Google, Facebook and many others.

There has been a growth in the usage of multi-factor authentication [Choudhury et al., 2011]. One-Time Passwords (OTP) [Khalid et al., 2013] is also a straightforward alternative to deploy a two-factor authentication mechanism. Federation, for instance, allows distinct entities to delegate authentication authorities and benefit from communications across different clients and platforms. Well known implementations of federation are Eduroam [Milinovic et al., 2008] or eduGAIN [López, 2006].

Other solutions include complete suites with several functionalities available. They pack many of those authentication systems along with additional functionalities and features, offering robust and more complete solutions. For instance, BlindIdM [Nuñez and Agudo, 2014] is a Identity as a Service (IDaaS) model with the particularity of being focused on providing privacy protection. Open Access Management (OpenAM) [OpenIdentityPlatform, 2020] is an access management solution that, among others, supports authentication, authorisation, SSO and Federation. Similarly, Keystone [OpenStack Foundation, 2020], AWS IAM [Amazon, 2012] or Microsoft Azure Active Directory (AD) [Chilberto et al., 2020] also provide such kind of services.

Some researchers claim for substituting passwords by cryptographic methods, like for example certificate-based or biometric solutions. However, deploying such systems is not a simple task. It imposes higher costs and leads to less usable software. After comparing several authentication solutions, it was concluded that password-based authentication was [Bonneau et al., 2012] and still is [Alqubaisi et al., 2020] the primary choice.

Additionally, in Cloud environments, supporting multi-tenancy at container level, providing QoS assurances, and flexible programming abstractions layers creates particular security challenges [Jasti et al., 2010; Chana and Singh, 2014]. Thus, it is necessary to design a strategy that allows achieving the levels of security required by Cloud-driven deployments.

3.1.3 Contributions

The contributions presented in this Chapter also considered the requirements of the H2020 EUBra-BIGSEA project. This project developed a framework, a platform and a library to ease the development of highly-scalable, privacy-aware data analytic applications running on top of Quality of Service Cloud infrastructures, reducing development cycles and deployment costs.

Considering the specific demands of the H2020 EUBra-BIGSEA project, which were the integration target of this solution, it was decided to initiate the implementation using simple yet effective solutions for each module. For instance, authentication is implemented using password-based mechanisms under end-to-end encryption. This allowed to establish an Application Programming Interface (API) architecture and to determine the best service interface.

In order to manage authorisation and accounting, a straightforward and transparent API was proposed and implemented. The API allows integration with external identity providers using technologies like OAuth.

Provisioning of AAA services should be elastic and efficient to cope with the requirements of Cloud systems. In addition to infrastructure and application support, configuration and deployment simplicity are the main features of the proposed AAAaaS solution.

The proposed AAAaaS solution, described next, provides general functionalities of traditional AAA and IAM services. Additionally, it is possible to interface it with external identity providers. The software is deployable and manageable according to three fundamental Cloud principles: scalability, elasticity and resilience.

3.2 Towards a Cloud-based AAA Mechanism

The proposed AAAaaS is a solution that is seamlessly deployed and integrated into the most diverse setups (from local development environments to multi-node clusters). It is envisioned to comply with cloud-based deployment and life cycle management strategies, which is not commonly available in related work. This offers flexibility, interoperability, and decentralisation of services, which are highly valuable for Cloud environments.

The requirements and architecture of the solution are described next.

3.2.1 Requirements and Architecture

The proposed solution provides traditional functionalities of AAA and IAM services. Additionally, it may interface with external identity providers. The solution is deployable and manageable according to three fundamental Cloud principles: scalability, elasticity, and resilience. Specifically, the solution meets the following requirements:

- Providing a Cloud manager framework-agnostic solution.

- Supporting Business to Consumer IAM functionalities.
- Supporting external identity providers.
- Supporting Access Control Management functionalities.
- Providing a common authentication to the infrastructure and applications.

Meeting the requirements mentioned above means the design must serve the needs of different users, infrastructures and applications. The system also needs to support different environments, from standalone pre-configured deployments to more advanced configuration setups such as Cloud computing clusters with multiple nodes and auto-scaling mechanisms.

Figure 3.1 presents the architecture of the proposed AAAaaS. It follows a modular approach where storage, back-end and front-end components form the core of the solution. Each component can be managed and scaled according to specific deployment restrictions and usage patterns. The web server is based on Nginx and the Representational State Transfer (REST) API is implemented using Pyramid [Tavares et al., 2010]. Nginx is necessary for load balancing and forwarding traffic to Pyramid instances. All internal components' communication is encrypted, and CloudFlare SSL [Cloudflare, Inc., 2020] is used to generate and manage certificates.

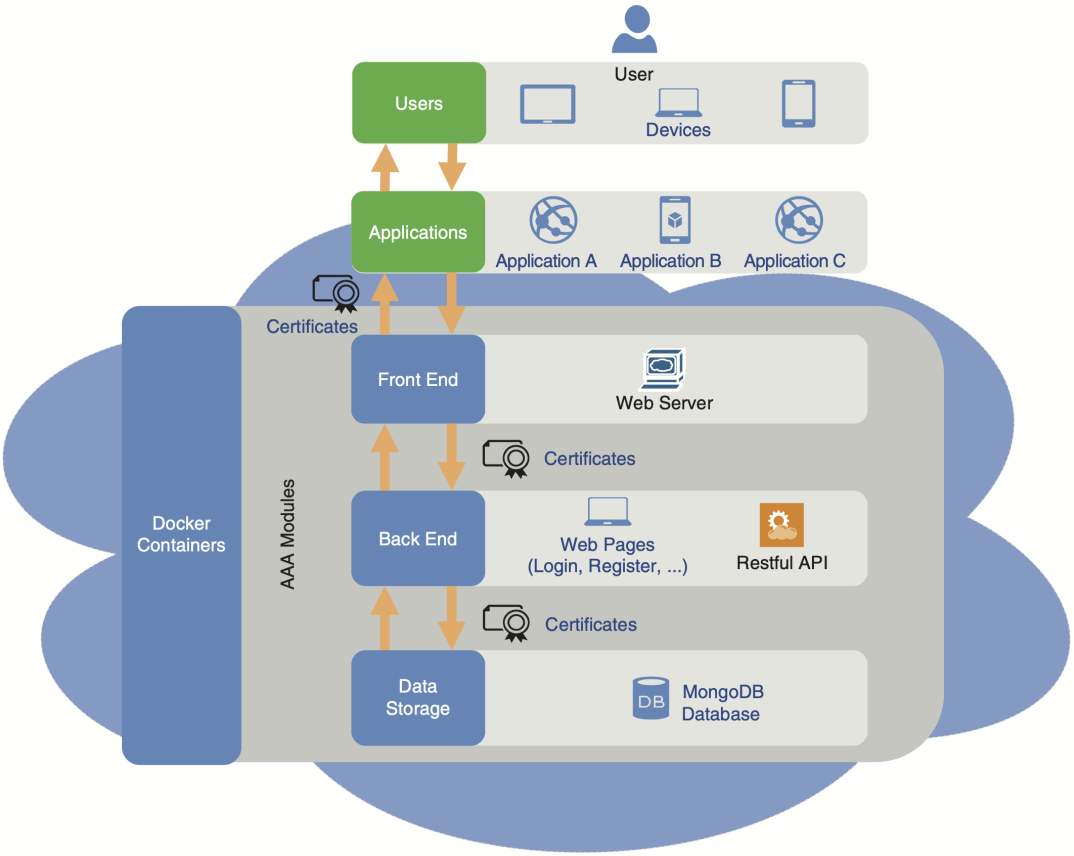


Figure 3.1: AAAaaS architecture – users and applications' interaction with AAA service in the Cloud.

The modular development not only complies with Cloud computing fundamentals but also helps developers fine-tune the system to their specific needs. By having separate components, it grants the possibility of adding or changing particular components. For instance, the database container can be replaced by others, more suitable to specific usage scenarios.

3.2.2 Internal Components and Implementation

The solution is designed in a containerised fashion, with direct relationships between different components (i.e., web server, application with a REST API, and database). Docker images containing the necessary code, settings, and dependencies are available on Docker Hub [1]. The images can be pulled from there, allowing the creation of containers with predefined or user-defined parameters. The components are described next.

Web Server

The AAAaaS Webserver is built with Nginx. Its role in this setup is to act as reverse proxy for the web application. All the requests directed to the web application go through the web server, that securely connects to the Web Application (described next) through SSL. The AAAaaS Webserver has been published as an open-source solution available on GitHub [2].

Web Application

The AAAaaS Webapp is a RESTful application that provides authentication, authorisation and accounting. The service is offered with a Graphical User Interface (GUI) and a REST API. It securely connects to the Web Server and Database through SSL.

The web application is developed in Python. Since it is a web service, it was developed with a suitable Python Web Framework: Pyramid. This framework is lightweight and offers clear documentation, customisation, performance and extensibility. The REST API of the service is deployed in this component. There are several endpoints (Section 3.2.3) available for end-users, administrators or services to access the AAAaaS Webapp. The access can be made through the REST API or directly through a GUI.

The most common interactions with the service are accessible through a GUI. A set of web pages provide functionalities such as sign up, sign in, modifying user information, and resetting the password. Nevertheless, all functionalities offered through the GUI are also available, among others, through a REST API. Examples of other functionalities available through the API are creating favourites, authorisation rules, and accounting. The service allows not only users but also applications and services to access its features. The AAAaaS Webapp has been published as an open-source solution available on GitHub [3].

[1] https://hub.docker.com/u/paulo308
[2] https://github.com/paulo308/AAAaaS_Webserver
[3] https://github.com/paulo308/AAAaaS_Webapp

Database

The AAAaaS Database is built with MongoDB. During the design phase, the possibilities were plentiful. Nevertheless, the choice was eventually made based on MongoDB flexibility as it is a cross-platform, Not only Structured Query Language (NoSQL), open-source and document-oriented. The information is stored using JavaScript Object Notation (JSON) documents. The Webapp securely accesses the database through SSL. The AAAaaS Database has been released as an open-source solution available on GitHub [4].

The API endpoints and respective AAAaaS functionalities are described next.

3.2.3 REST API Endpoints and Functionalities

Table 3.1 describes the authentication requests that can be made to AAAaaS. The first column of the table describes the actions, and the second shows the calls' addresses. The methods enable the following functionalities: signing up of a new user; signing in of an existing user; verification of a token associated to a user; verification of token and retrieval of complete user information; signing out of a user; updating/changing user information; deleting user account; changing user password; and retrieval of a forgotten password.

Table 3.1: Authentication – REST API endpoints.

Action	**Address**
Verifies credentials and provides user data	engine/api/checkin_data
Verifies token and replies with username	engine/api/verify_token
Verifies token and replies with detailed user information	engine/api/read_user_info
Creates user entry in database	engine/api/signup_data
Invalidates token from user	engine/api/checkout_data
Updates user information	engine/api/update_user
Deletes user account	engine/api/delete_user
Changes the password	engine/api/change_password
Sends email with new password	engine/api/forgot_password

Table 3.2 describes the authorisation requests that can be made to AAAaaS. The first column of the table describes the actions, and the second shows the calls' addresses. The methods enable the following actions: create authorisation, update authorisation, read authorisation, read authorisations, delete authorisation and update resource usage. The last row shows the endpoint that enables reading accounting information.

Table 3.3 describes the favourites' calls that can be made to AAAaaS. The first column of the table describes the actions, and the second shows the calls'

[4]https://github.com/paulo308/AAAaaS_Mongodb

Table 3.2: Authorisation and accounting – REST API endpoints.

Action	**Address**
Creates a new authorisation rule	engine/api/create_authorisation
Updates the authorisation rule	engine/api/update_authorisation
Reads authorisation rule	engine/api/read_authorisation
Reads all authorisation rules	engine/api/read_authorisations
Deletes authorisation rule	engine/api/delete_authorisation
Updates resource usage	engine/api/use_resource
Reads accounting information of user	engine/api/read_accounting

addresses. The methods enable the following actions: create favourite, retrieves favourite, retrieves all favourites, and deleting favourite.

Table 3.3: Favourites – REST API endpoints.

Action	**Address**
Creates new user favourite	engine/api/create_favorite
Reads user favourite	engine/api/read_favorite
Reads all user favourites	engine/api/read_favorites
Deletes user favourite	engine/api/delete_favorite

Table 3.4 describes the email association calls that can be made to AAAaaS. The first column of the table describes the actions, and the second shows the calls' addresses. The methods enable the following actions: create email association, retrieves associated emails, and deleting associated email.

Table 3.4: Email association – REST API endpoints.

Action	**Address**
Creates new email association	engine/api/create_email
Reads associated emails	engine/api/read_emails
Deletes associated email	engine/api/delete_email

The following section describes how AAAaaS is integrated in the intercontinental EUBra-BIGSEA Cloud platform.

3.3 Integration and Validation in the Scope of the EUBra-BIGSEA Cloud Platform

The EUBra–BIGSEA Cloud platform [Silva et al., 2018; Alic et al., 2019] is divided into three main layers: Applications, Data, and Infrastructure. Security and privacy mechanisms are orthogonal to the three layers. Therefore, the security and privacy mechanisms provide authentication, authorisation, and accounting (through AAAaaS), data privacy protection and trustworthiness estimation. Figure 3.2 shows EUBra-BIGSEA architecture layers and how the different components interact. AAAaaS is visible on the top left corner.

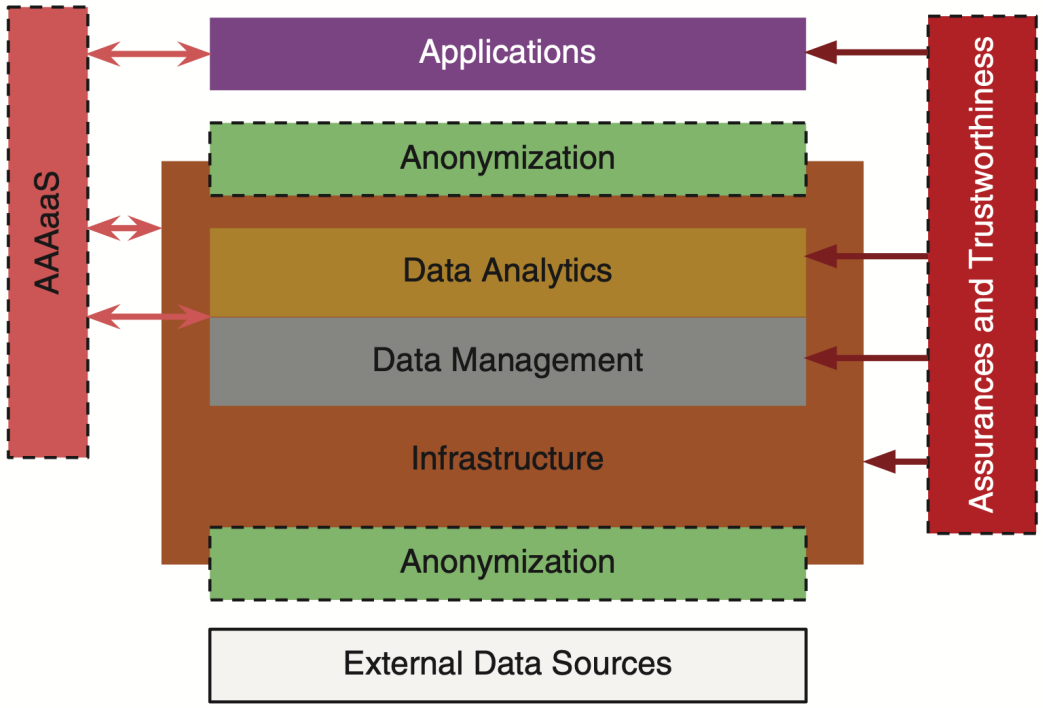


Figure 3.2: EUBra-BIGSEA Cloud platform architecture [Silva et al., 2018].

The platform uses Mesos [Heidari et al., 2016] and Marathon [Mesosphere, Inc., 2020] to orchestrate the deployed services and AAAaaS is fully compatible with those mechanisms. Moreover, AAAaaS was adopted and integrated by services and application such as *Routes 4 People* (web application), *Melhor Busão* (Android App), *Municipality Dashboard* (public service) and *Ophidia Server*, an infrastructure management service.

In terms of usability and functionality, some applications or services rely more on authentication features such as sign-in, token verification, or change of user information. In contrast, others tend to use more authorisation features like resource allowance verification. Nevertheless, all of the options are available at all times. During the stress testing campaigns that were performed, AAAaaS was able to maintain the throughput as the number of parallel clients increased and had a failure rate inferior to 0.5%. Moreover, internal testing also showed that end-to-end encryption was ensured at all times with user communication as well as component communication.

Based on feedback from the partners involved in the integration of the service, AAAaaS is intuitive to set up and integrate. It requires the consultation of the REST API documentation to verify the endpoints' parameters and just a few lines of code to complete the integration. The feedback was similar when the GUI is used, needing only an additional block of code to redirect the user to the service and to handle the responses after the user interaction. In this case, the interaction between parties is seamless for the users, as they are automatically redirected and all operations are transparent.

3.4 Summary

This chapter proposed AAAaaS, a cloud-based security solution that was required to integrate with other components and serve different applications and services from different abstraction layers. Therefore, it was crucial to start with a simple but effective API, designed to fit the specific needs of other partners and, at the same, time respect important architectural decisions. The AAA library evolved to the solution proposed in this chapter. It offers both a web interface and REST API, which have been deployed and successfully used in the EUBra-BIGSEA Cloud platform and many of its Use Case applications.

Overall, the result is a differentiated Cloud service that, on the one hand, offers diverse compatibility, ease of use, and seamless integration and, on the other hand, provides security, scalability, resilience, and flexibility.

The outcomes of this chapter resulted in the following publications:

- Alic A., et al. (2019). **BIGSEA: A Big Data Analytics Platform for Public Transportation Information**. Future Generation Computer Systems, 96, 243 - 269. DOI: 10.1016/j.future.2019.02.011
- Silva P. et al. (2018). **A Europe-Brazil Context for Secure Data Analytics in the Cloud** in IEEE Security & Privacy, vol. 16, no. 6, pp. 52-60. DOI: 10.1109/MSEC.2018.2875326.
- Alic A. et al. (2018). **GIS and Data: Three Applications to Enhance Mobility**. In XIX Brazilian Symposium on Geoinformatics - GeoInfo 2018, Campina Grande, PB, Brazil, December 5-7, (pp. 1–12).

The following chapter presents data analysis mechanisms that increase privacy preservation through the identification of PII, contributing to personal data protection and increased privacy preservation in Cloud environments.

Chapter 4

Data Analysis for Privacy Preservation

Contents

Information systems and services handle a plethora of data types. As such, mechanisms that help organisations protecting the involved data subjects are of the utmost importance. In this chapter, different aspects of data analysis (i.e., tool-based, model-based, and hybrid analysis) are explored in order to propose privacy-preserving data analysis mechanisms.

The outcomes of this chapter resulted in the following publications:

- Silva, P. et al. (2020). **Risk Management and Privacy Violation Detection in the PoSeID-on Data Privacy Platform**. SN COMPUT. SCI. 1, 188. DOI: 10.1007/s42979-020-00198-9.
- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2020). **Using NLP and Machine Learning to Detect Data Privacy Violations**. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, pp. 972-977. DOI: 10.1109/INFOCOMWKSHPS50562.2020.9162683.
- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2020). **Using Natural Language Processing to Detect Privacy Violations in Online Contracts**. In Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC '20). Association for Computing Machinery, New York, NY, USA, 1305–1307. DOI: 10.1145/3341105.3375774.
- Casaleiro, R. et al. (2020). **Protection and Control of Personally Identifiable Information: The PoSeID-on Approach**. Journal of Data Protection & Privacy, 3(2).

4.1 Introduction

Data analysis is performed for several reasons, ranging from data validation and compliance to value extraction and classification. The data types available can be structured (e.g., tables), unstructured (e.g., documents or text), with numerical or textual data. Among many other possibilities, the analysis also enables businesses to make decisions or to offer a variety of services. Many times, these services are provided free of cost in exchange for valuable user data, which in most cases is Personally Identifiable Information (PII). Since data can be sensitive and private, it is essential to devise and employ privacy-preserving PII analysis mechanisms.

This chapter proposes mechanisms that allow the automation of data analysis while maintaining the characteristics of a privacy-preserving system. Such mechanisms include AI-based approaches such as NLP tools (Section 4.3) and other ML mechanisms (Section 4.4). The mechanisms were integrated and validated in the scope of the H2020 PoSeID-on Project. Furthermore, there is a discussion of the lessons learned and how this approach can effectively act as a Privacy Enhancing Technology.

4.1.1 Background

NLP tools are software libraries and applications that are used for extracting information in digital format and derive meaning from it. The analysis can be performed, for instance, over semantics, syntax, or speech, with each one having its particular challenges. Most tools support tokenisation, Part of Speech (POS) Tagging or lemmatisation (i.e., identifying the lemma – the dictionary form). Other features such as Named Entity Recognition (NER), translation, recognising textual entailment or Natural Language Understanding (NLU) are more specific.

Among others, these tools can be used for spam detection [Jindal and Liu, 2007], fraud detection [Ngai et al., 2011] or, generally, document classification and analysis. NLP is a subset of AI. Figure 4.1 shows how NLP relates to ML and Deep Learning (DL). Several ML applications are based on supervised learning approaches [Ayodele, 2010], as is NLP.

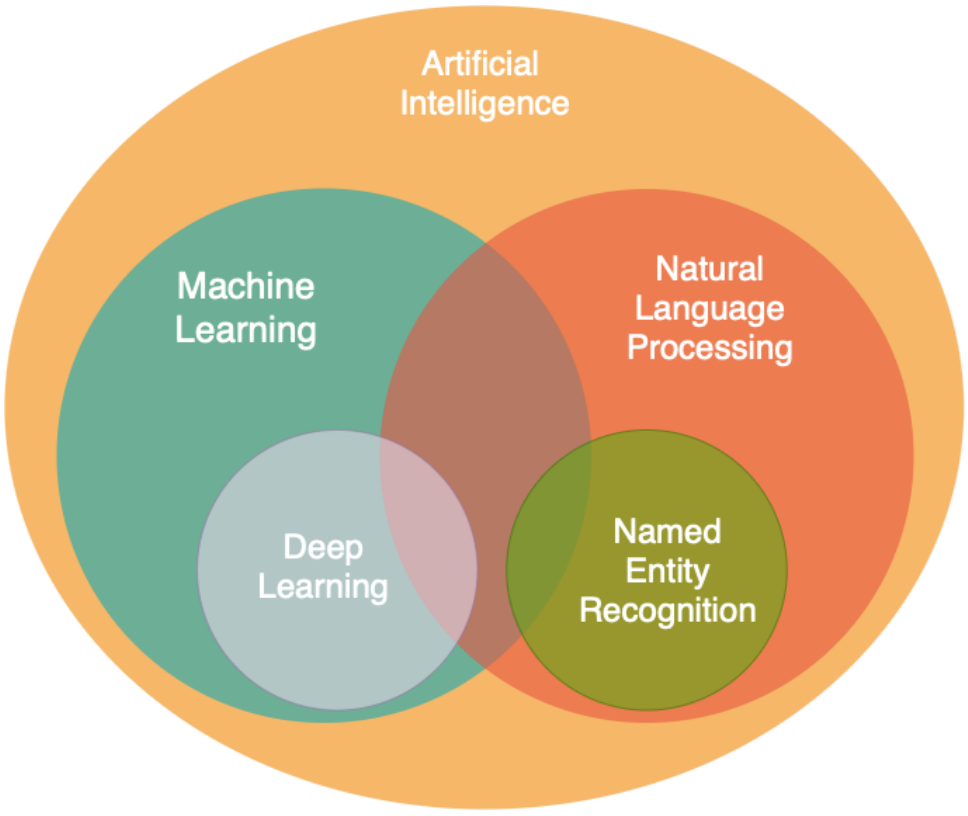


Figure 4.1: NLP relationship with AI (adapted from [Athena Tech, 2019]).

NER, one of NLP's sub-tasks, finds and classifies named entities according to a set of categories [Yadav and Bethard, 2018]. Those categories can be people's names, addresses, states, countries, money, organisations, laws or any other kind of PII. With NER it is possible to automatically scan text documents, data structures (or any other text file container) and understand the importance of those entities in the context of the text. Performing NER with different NLP tools may lead to different NER performances due to various machine learning implementations. Also, a NER system designed within a tool for one project

may execute differently in another project or not do the task at all [Ratinov and Roth, 2009].

Three of the most representative NLP tools, Natural Language Toolkit (NLTK), Stanford CoreNLP and spaCy, have been used in part of the experimental tests described in Section 4.3. Those tools, described next, are extensively used by the community and have a proven track record.

NLTK [Bird et al., 2009] is one of the most well-known NLP tools. It is community-driven and open-source Python software, which allows the manipulation of different corpora, categorising text or analysing linguistic structure. It implements a Naive Bayes classifier [Chen et al., 2009] and Hidden Markov Models (HMM) [Lafferty et al., 2001].

Stanford CoreNLP [Manning et al., 2014] stands out as a reference tool in the field of NLP. It is open-source and developed in Java. Among other features, it is capable of performing sentiment analysis, dependency parsing, or NER, for instance. Stanford CoreNLP applies Conditional Random Fields (CRFs) – probabilistic models able to perform segmentation and labelling of sequential data [Lafferty et al., 2001] such as the text used in NLP tasks.

ExplosionAI introduced SpaCy [ExplosionAI, 2020], claimed to be the fastest NLP library in the world. The fact is that it is not only fast, but also performs well against similar tools and supports similar features. Spacy uses Convolutional Neural Networks (CNNs) [Zhang and Suganthan, 2016] with pre-trained word vectors [Pennington et al., 2014] to train its models.

Word vectors are useful to analyse relationships across words. They are vectors of numbers that represent the meaning of words. Multidimensional continuous floating-point numbers represent the words, and semantically similar words are mapped to proximate points in geometric space [Pennington et al., 2014]. In simpler terms, a word vector is a row of real-valued numbers where each point captures a dimension of the words' meaning and where semantically similar words have similar vectors. SpaCy uses pre-trained word vectors in some of its models. The approach described in Section 4.4 uses them as well.

The previously described mechanisms are very useful for a privacy-preserving data analysis when the scope of analysis is PII. As PII is sensitive and private, it is essential to devise and employ privacy-preserving text analysis mechanisms like the ones proposed in this Chapter.

4.1.2 Related Work and Open Issues

The literature includes extensive work and publications regarding NLP, its characteristics, and its performance. For instance, Omran and Treude [Al Omran and Treude, 2017] perform a systematic literature review on how to choose an NLP library. The most commonly mentioned NLP tools are the NLTK, Stanford CoreNLP and spaCy.

The NLP's sub-task more suitable for the type of analysis described in this thesis is NER, as this mechanism uses models to classify entities (e.g., Persons

or Locations) it finds in text. Jiang et al. [Jiang et al., 2016] reviewed several tools, to assess which ones are more accurate in NER. Of course, this can be applied in a wide variety of fields. For instance, Ritter et al. [Ritter et al., 2011] used NER to recognise Named Entities in tweets, and Vlachos [Vlachos, 2007] evaluated NER systems for biomedical data.

Other noteworthy NLP tools include: TextBlob [Loria, 2019]; Polyglot [Al-Rfou et al., 2015], developed in Python and inspired in NLTK; the General Architecture for Text Engineering (GATE) [Cunningham et al., 2000], which comprises several components for specific purposes such as Hypertext Markup Language (HTML), Extensible Markup Language (XML) or email processing; and Google's SyntaxNet [Google, 2019], which uses the open-source TensorFlow [TensorFlow, 2019] machine learning platform to provide an NLU toolkit.

Words can have different meanings in different contexts (i.e., polysemy), which can be a limitation for some NLP mechanisms. To tackle this problem, Peters et al. [Peters et al., 2018] proposed Embeddings from Language Models (ELMo), to allow a word to have multiple embeddings, depending on the context. Shortly after, Howard et al. [Howard and Ruder, 2018] proposed Universal Language Model Fine-tuning for Text Classification (ULMFiT), an enhanced approach able to provide similar results with less training data.

NLP's state of the art is currently based on models proposed by research teams at Google – the Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] – and Facebook – the Robustly Optimized BERT Pretraining Approach (RoBERTa) [Liu et al., 2019]. BERT is a bi-directional transformer model for pre-training over a lot of unlabelled textual data to learn a language representation and is currently used in Google's search engine. RoBERTa is a similar model with slightly increased performance.

Another way of performing data analysis is by building an ensemble of classifiers (i.e., a combination of classifiers). The purpose of an ensemble mechanism is to aggregate multiple models so that the final prediction is improved [Rokach, 2010]. Traditional approaches typically rely on weighting methods such as majority voting, where the final prediction is the entity that obtained the highest number of votes. The mechanism proposed in Section 4.5 is based on a similar approach.

Regardless of the recent advances, there are still open issues related to current data analysis mechanisms, such as a lack of research driven by privacy-preserving data analysis. Most data analysis outcomes have the goal of providing valuable user information so that business can profit from targeted advertising and related approaches but not for the benefit of the user and his/her privacy.

Despite the availability of comprehensive NLP research in the literature, there is still insufficient work relating NER, PII, and privacy implications. There is a high focus on clinical or biomedical data but not in the broad spectrum of PII, which encompasses different kinds of personal information. Since PII is a sensitive type of data and it is not openly accessible, it is not trivial to acquire the most suitable datasets.

Although most of the available datasets (Table 4.1) feature entity annotation, many are not publicly accessible. As such, in this work, preference was given to datasets that are publicly accessible and contain labelled entities. The available information describes the datasets as mostly machine-annotated, requiring human intervention for correction or validation. Overall, the limited data, its scope and annotation limitations make it a challenge to train ML models.

As explained before, application of the aforementioned mechanisms is often used for data analysis. Additionally, their application is also usually a part of Extract Transform Load (ETL) processes or pre-anonymization steps. Moreover, several approaches are often offered as a service [Canale et al., 2018]. However, although very capable, the issue with these approaches is that they may not offer as much trustworthiness or privacy assurances as other counterparts.

Overall, there is a lack of solutions or services where the employment of these mechanisms can directly benefit data owners. The next section shows how the mechanisms proposed in this Chapter are able to provide a privacy-preserving data analysis.

4.1.3 Contributions

As long as there is transparency and compliance with regulations, using AI through NLP and NER models can be a strong Privacy Enhancing Technology. Applied in privacy-preserving data analysis (e.g., active or passive monitoring of text for compliance verification), it largely avoids the involvement of human operators or data retention mechanisms.

In order to increase privacy assurances, it is necessary to design mechanisms that can not only provide privacy assurances but also increase automation and reliability. NER models are the ideal candidates as a mean to monitor data and detect privacy violations.

The contributions of this chapter are presented in the next sections.

Section 4.2 provides a description of the methodology, datasets and evaluation metrics. Section 4.3 discusses how to generate purposeful training and testing datasets based on publicly available data. Section 4.4, presents NER models specifically created with ML algorithms such as MLP [Ruck et al., 1990] and RF [Zhang and Ma, 2012]. Finally, Section 4.5, presents a privacy-preserving hybrid NER pipeline for PII analysis and identification. The following section describes the approach followed to propose a privacy-enhancing data analysis mechanism.

4.2 Methodology

This section describes the approach followed in the study, development and validation of ML mechanisms for the identification, monitoring and validation of PII in a privacy-preserving manner. The evaluation metrics and data are also described.

4.2.1 Overall Approach

The very first step to propose an intelligent and automatic data analysis mechanism was to experiment with previously existent and representative NLP tools. It was necessary to devise a methodology to help determine to which extent NLP and NER can reliably detect and identify PII and, ultimately, be used as a Privacy Enhancing Technology (PET). For that purpose, different tools are used to train (and test) NER models, with different datasets.

This approach allows the evaluation of not only the NER models but also the NLP tools used: NLTK (V3.4), Stanford CoreNLP (V3.9.2), and spaCy (V2). Moreover, the default $F_1 scores$ of each English NER model available with the tools is 0.85 [Bird et al., 2009], 0.86 [Finkel et al., 2005], and 0.85 [Explosion AI, 2020]. These values will be useful later on, to compare and analyse the mechanisms proposed in Sections 4.4 and 4.5.

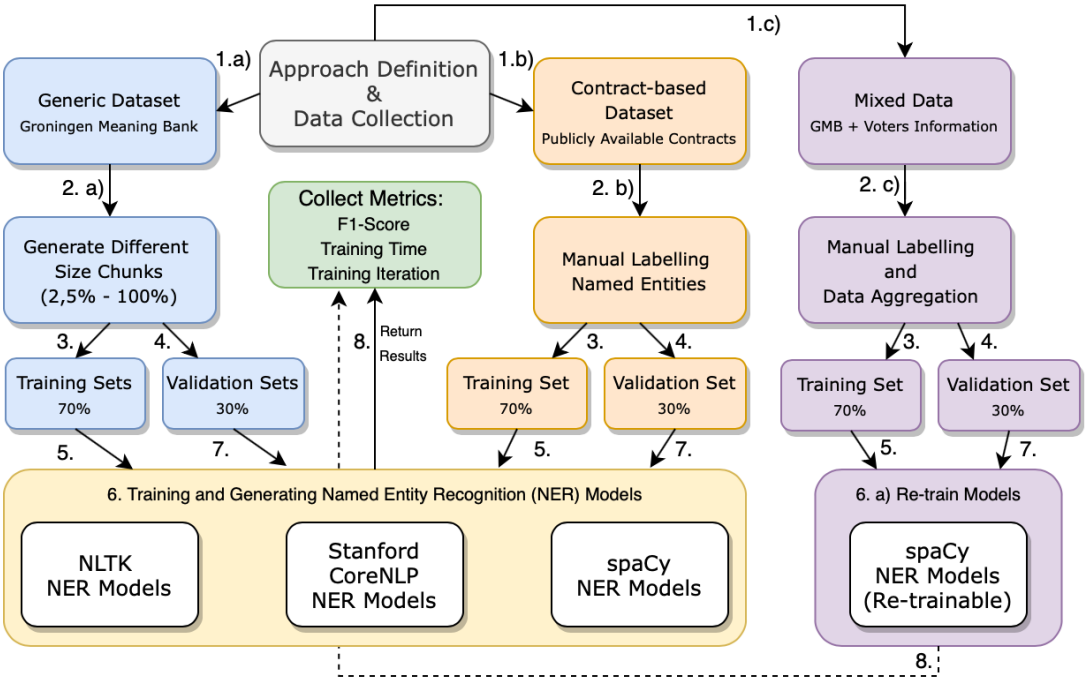


Figure 4.2: Model training approach.

Figure 4.2 shows every step of the process. After collecting the necessary data in steps 1.a), 1.b) and 1.c), it was necessary to partition the data in different chunks to study data size (step 2.a)) and manually label the data (steps 2.b) and 2.c)). From that point on, the approach is identical: training and validating models in each one of the NLP tools (steps 3, 4, 5, 6 and 7). The only difference, in this case, is that with spaCy, an additional batch of tests was performed to analyse the model re-training capabilities.

By adopting this approach, it became possible to analyse the following aspects:

- *Content influence* – Model training with different contents. The aim is to discover how is the classification performance affected by a generic and

a content-specific dataset. More specifically, as detailed in Section 4.2.2, the generic dataset was based on news reports and the context-specific dataset was based on voters' information and publicly available contracts. Moreover, a combination of these two kinds was also used for the evaluation.

- *Size influence* – ML models tend to perform better when trained with large and/or diversified amounts of data. Partitioning the original datasets into smaller portions enabled the analysis of the dataset size influence. The partitioning required a reduction the datasets size: from 2.5% to 100% of their original sizes (as explained in Figure 4.2 and shown in Section 4.3.1).
- *Classification performance* – One of the most important aspects to account for is the classification performance (i.e., accuracy, precision and $F_1 scores$). This determines how well the data is classified. Additional details of the metrics are provided in Section 4.2.3.
- *Training performance* – In this case, it is possible to measure how long the training process takes. One aspect that influences model training time is the number of iterations (i.e., epochs).

Afterwards, different ML algorithms are used for the training of NER models (cf. Section 4.4. Since Multilayer Perceptron (MLP) [Ruck et al., 1990] and Random Forests (RF) [Zhang and Ma, 2012] were the algorithms applied in the model training, it was necessary to use Word Vectors to convert text to numerical representation. For that, three different word vectors approaches were used: two of the most representative pre-trained word vectors (Google and gloVe), and a manually trained word vector. Having distinct approaches allows for a more comprehensive analysis and comparison. The three approaches were the following:

- *Training own word vectors* – Using Word2vec [Google Code Archive, 2020] to create own word vectors from training data described in Section 4.2.2. Although it does not result in a representative word vectors, it is relevant for the sake of analysis.
- *Google News pre-trained word vectors* – This approach includes 3 million words and phrases trained on roughly 100 billion words from a Google News dataset [Mikolov et al., 2013; Google Code Archive, 2020]. Each vector has 300 features.
- *GloVe pre-trained word vectors* – GloVe word vectors are trained on Wikipedia 2014 and Gigaword5 datasets, which means being trained on 6 billion words [Pennington et al., 2014]. Each vector also has 300 features.

For the datasets used in the experimental work (described in Section 4.2.2 and Table 4.1), it was necessary to divide the original dataset into two, for training and validation purposes, respectively. The first part, for training, contained 70% of the dataset, while the second part, for validation, contained the remaining 30%. These proportions are commonly used across the AI field. After preprocessing the data, it was possible to proceed to model training and evaluation

of each tool. Except for the number of training iterations, the default settings were kept unchanged throughout the entire process.

After training the model, it is necessary to evaluate the models' performance. This is, assess how well it predicts entities using datasets that it has never processed before. For that, it is necessary to provide each model with validation datasets which are equally labelled, as the training datasets. Each model then classifies the entities in the validation dataset and then compares the results to the actual label that corresponds to each entity.

The results are analysed in Section 4.3 and 4.4. The findings gathered from these steps allowed for an improved approach with the a hybrid classification mechanism (proposed in Section 4.5). The next section describes the types of data that are used to train the models.

4.2.2 Data and Named Entities

Labelled datasets are particularly useful for training models in capable of data analysis tasks. Their usual denomination is *corpus*: a text collection. The labelling can be manual (i.e., gold standard) or automated, and it contains a tag for each word. The context of a word in a sentence makes a difference. For instance, "The little yellow cat" is semantically represented as "The/DT little/JJ yellow/JJ cat/NN", where DT stands for Determiner, JJ for Adjective, and NN for a Noun.

Data can be also be labelled (i.e., annotated) on a variety of topics such as cultural, financial, political, scientific and many others. For instance, in the previous example, "cat" would have the label "animal". This aspect is what matters the most in the scope of the mechanism proposed in this chapter. Despite the availability of several datasets, performing data analysis on PII is particularly challenging since it is a sensitive type of data, and it is not easily available for obvious reasons.

To circumvent the unavailability issue, at least two options can be considered. One is to generate synthetic data that resembles real data as closely as possible. For that end, solutions such as MostlyAI [MostlyAI, 2020] can be of use. Another option is to retrieve publicly available information with the necessary entities. In this way, it is possible to train the proposed models to identify PII with common entities such as names, addresses, locations, events, and others. There are other methods like regular expressions that are also used to identify a variety of entities such as phone number, email, credit card number and many others.

Table 4.1 shows available datasets that may be used for NER model training. It is possible to observe that some fields are not completed as it was not possible to obtain information such as entities or the overall size of the dataset. SpaCy is released with models trained on OntoNotes5, which is one of the most extensive and most annotated datasets. On the other hand, NLTK is released with models trained on Brown dataset and Stanford CoreNLP on CoNNL 2003 and MUC 6/7.

Table 4.1: Datasets characteristics.

Dataset	**Named Entities**	**Public Availability**	**Size**
OntoNote5	✓	for a fee	1.445.000 tokens
RCV1	✗	license agreement	810.000 new stories
CoNNL 2003	✓	license agreement	301.418 tokens
MUC 3 & 4	✓	✓	16755 unique tokens
MUC 6 & 7	✓	for a fee	Unknown
ACE 2002	✓	for a fee	Unknown
New York Times	✓	for a fee	1.8 million articles
GMB	✓	✓	1.374.629 tokens
GMB-derived	✓	✓	1.354.149 tokens
Enron email	✓	✓	200.399 messages
English Gigawork	✓	for a fee	1.756.54 tokens
Brown	✗	✓	1.000.000 tokens
Voters	✗	✓	112.351 voters
Metrolink and DIR	✗	✓	19836 tokens

The dataset used to validate the proposed mechanism is the Groningen Meaning Bank (GMB)-derived with 1.354.149 tokens. It was retrieved from Kaggle [Kaggle Inc., 2020], and although it is based on GMB data [University of Groningen, 2019], it is improved by the community – this was one of the reasons to choose the dataset, along with the type of named entities included in the dataset, as described bellow. They are mainly composed of public domain English text like news, reports and other publications.

The entities used in the Tool-based PII classification (Section 4.3) were the ones available in the datasets. Although not all are useful for PII-related analysis (e.g., artifact, event, and natural phenomenon), they were included for consistency purposes. The following entities were adopted:

- Geographical entity
- Organisation
- Person
- Geopolitical entity
- Time indicator
- Artifact
- Event
- Natural phenomenon

The entities used in the tool-based classification are somewhat limited. Therefore, to enhance the reach of the proposed mechanism, in the model-based classification (Section 4.4) the focus was on different, and additional, entities. For that reason, it was necessary to retrieve legal and publicly available datasets containing these kinds of entities (described next).

In this case, the datasets were context-specific, with an emphasis on content that can be considered as PII. Contracts are a top candidate to source these datasets, as they usually include detailed PII collective of individual entities. Since there

were restrictions regarding data sources, the choices resulted from online searches for publicly released contracts. The first source was a set of Metrolink and Texas Department of Information Resources (DIR) contracts [Metrolink, 2019; Texas DoIR, 2019] that was publicly and lawfully released. The second was the U.S Department of Defense (DoD) [U.S. DoD, 2019] daily contracts about the expenses of each military division. The third dataset contained 112.351 U.S.A. voters' registration data [N.C. Board of Elections, 2020]. The considered entities were the following:

- Person*
- Address*
- Organisation*
- City*
- State*
- Money*
- Country*
- Percent*
- Law*
- Bank details
- Employm.* [1]
- Title*
- Email*
- Post code*
- Time*
- Date*
- Date of birth
- Passport number
- Social sec. [2]
- License plate n.[3]
- Gender
- National ID n. [4]

The entities marked with an asterisk (*) were manually labelled. It is possible to observe that, from the total list of entities, 68% of them were labelled during the annotation process.

The retrieved datasets were in their original release formats. Therefore, additional data manipulation and transformation was necessary. Metrolink and DoD [U.S. DoD, 2019] contracts were manually labelled, since they were non-labelled datasets. The voters' dataset did not require the same steps, since it had a header description for each attribute - a custom script was sufficient to label the dataset.

The final step before proceeding to the training and validation sessions was to split the voters' dataset and merge with samples from other datasets. This allowed the study of the models' performance with different data contexts during training and validation. For instance, one sample of 1630 voters and 50k lines of the Kaggle results in a combined dataset. From now on, these samples will be denominated as *combined* data. The next section describes the metrics used during the experimental work.

[1] Employment contract & salary information
[2] Social Security Number
[3] License plate number
[4] National ID number

4.2.3 Evaluation Metrics

In order to analyse and evaluate the characteristics of NLP tools and the proposed NER models, it is necessary to identify the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These four elements are essential as they enable the measurement of the accuracy, precision, recall and $F_1 score$ of the models.

Accuracy is the ratio between the number of correct choices to the total number of choices. This metric should be applied when there is an even class distribution. Therefore, there might be cases where it is not provided, as it does not provide the desired insight.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

Precision is the ratio between the number of correct positive choices to the total number of positive choices.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Recall is the ratio between the number of positive choices to the real number of choices. This means not only true positives but also false negatives (i.e., false negatives are positives).

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$F_1 score$ is the weighted average between precision and recall. Contrary to accuracy, this metric becomes more helpful when dealing with uneven class distributions, which is the expected case in the Personal Data Analyser. It is used in related work to evaluate the models built for different purposes.

$$F_1 score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4.4}$$

Although not relevant for the analysis performed in the scope of this chapter, the time required to train each model, as well as the number of iterations (i.e., epochs) were also registered. Since the tool-based model training was performed in different environments (due to logistic and operational constraints), these measurements are merely indicative. Moreover, due to memory constraints, the number of iterations was limited to 500 in Stanford CoreNLP and spaCy. NLTK did not support the possibility of setting a specific number of iterations.

The performance analysis of the proposed mechanisms (as described in Section 4.5) when deployed in Cloud environments is also important. The metrics used to evaluate the performance in such environments were *throughput* (i.e., number

of analysis performed per second) and *average processing time* (i.e., the average time necessary to perform one analysis).

The following section provides a summary and analysis of the results obtained.

4.3 Tool-based Classification

As stated in the previous sections, the effectiveness and performance of the NLP tools was first evaluated in a generic dataset. Then, the tools are fed with datasets based on content-specific data such as contracts or voters' registration data.

The steps followed, as well as results, are described next. The results show a generally positive performance in accurately classifying entities both in generic and context-specific data.

4.3.1 Evaluation with Generic Data

To better evaluate the performance of the tools and respective models, the dataset was partitioned in smaller chunks. The objective was to assess how the performance of the models was affected by the dataset's size. The dataset was sliced in smaller portions (2.5%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%). For each portion, the 70% and 30% proportion rule was applied, for training and validation, respectively.

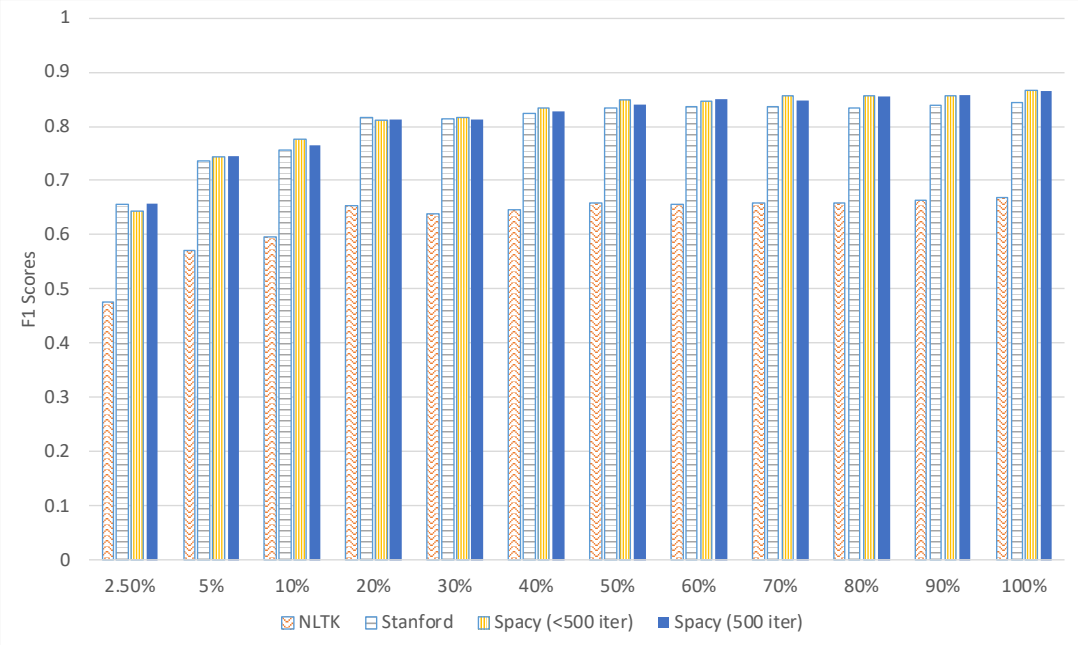


Figure 4.3: $F_1 scores$ (NLTK, Stanford CoreNLP and spaCy).

Figure 4.3 shows aggregated results of the NLP tools with generic data partitioned in different sizes. NLTK obtained a $F_1 score$ of 0.47 using the smallest

portion of the dataset (2.5%). Stanford CoreNLP and spaCy reached approximately 0.65.

Using the entire dataset (100%) provided the best results: NLTK achieved approximately 0.67, while Stanford CoreNLP and spaCy obtained 0.84 and 0.86, respectively. There was no significant difference between the 20%-sized dataset and the larger ones. The F_1score difference between the 20%-sized dataset and the full dataset is between 0.03 and 0.05, between Stanford CoreNLP and spaCy.

The lowest classification performance was observed in NLTK. On the other hand, Stanford CoreNLP and spaCy achieved similar results, with spaCy performing slightly better. The results indicate that, without any tuning of the model training settings, spaCy provides the best results for F_1score on a generic dataset. Additionally, training the models in spaCy for less than 500 iterations (i.e., different values under 500) provides similar results and requires much less training time (as shown in Figure 4.3, label "Spacy (<500 iter)"). This particular test was performed only to determine if a smaller number of iterations would influence the training time.

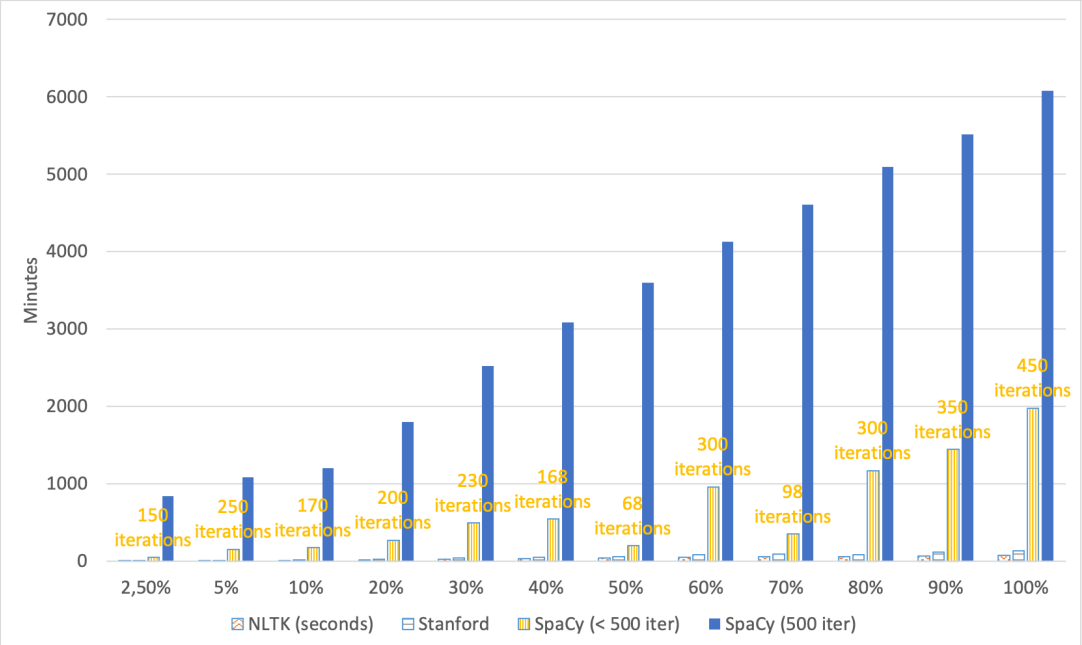


Figure 4.4: Model training times (NLTK, Stanford CoreNLP and spaCy).

Regarding training time (Figure 4.4), NLTK was the fastest. The elapsed time was approximately 2 seconds for the smallest dataset and approximately 75 seconds for the largest dataset. Stanford CoreNLP takes approximately 10 minutes to train the smallest dataset. On the other hand, it takes approximately 120 minutes to train the largest dataset. The training time with spaCy differs according to the number of training iterations. Setting a maximum of 500 iterations, the training time for the largest dataset was close to 6000 minutes. However, when the number of iterations is reduced to half (as shown in Figure 4.4, label "SpaCy (<500 iter)"), the same dataset size takes approximately 2000

minutes and the $F_1 score$ is very similar. Therefore, it is possible to achieve good results while spending less time training. SpaCy takes longer periods to train its models due to the underlying Neural Network. On the other hand, NLTK with Hidden Markov Models (HMMs) and Stanford CoreNLP with Conditional Random Fields (CRFs) are much faster.

4.3.2 Evaluation with Context-specific Data

The procedure for the context-specific data (i.e., the previously described contracts) was identical to the generic data approach. The dataset created was a combination of publicly released contracts [Texas DoIR, 2019; Metrolink, 2019] and contracts from the U.S. DoD [U.S. DoD, 2019].

After retrieving publicly available contracts in Portable Document Format (PDF), it was necessary to extract the information and convert it to text files. Only then was possible to perform the required tokenisation and proceed with the manual tagging of entities.

The focus was on different entities, namely those mentioned in Section 4.2.2. One of the sources was the U.S. DoD, that publishes the daily expenses of the military branches [U.S. DoD, 2019]. The other sources were publicly available contracts released by other entities [Texas DoIR, 2019; Metrolink, 2019]. It is possible to observe that, from the list of entities defined above, 68% of them were manually labelled during the annotation process.

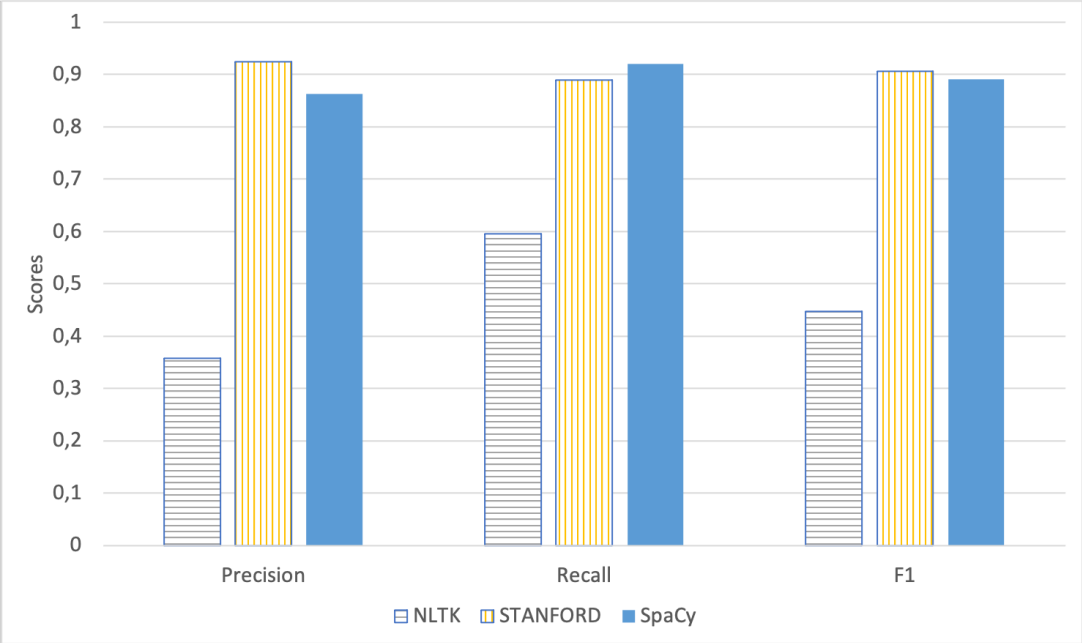


Figure 4.5: Precision, Recall, and $F_1 scores$ of models trained with context-specific data (NLTK, Stanford CoreNLP and spaCy).

Figure 4.5 shows the precision, recall, and $F_1 scores$ obtained while evaluating the models created from manually-labelled contracts. It was possible to observe that NLTK's highest scores were approximately 0.45. On the other hand, Stanford

CoreNLP and spaCy reached very similar scores (approximately 0.90). The difference between these two is 0.01, being Stanford CoreNLP the one with higher score.

4.3.3 Evaluation with Combined Data

To address the generalisation capabilities of the models, combinations of different types of data were created and used for training and validation under several circumstances. One of the cases is training models with generic data (e.g., Kaggle) and validating with context-specific datasets (e.g., U.S DoD contracts or United States (US) voters' registration data). Additionally, in some cases, US voters' registration data [N.C. Board of Elections, 2020] was used in order to increase the diversity of the dataset. Part of the approach included re-training models. Since NLTK and Stanford CoreNLP do not support re-training, spaCy was the only tool used in this scenario.

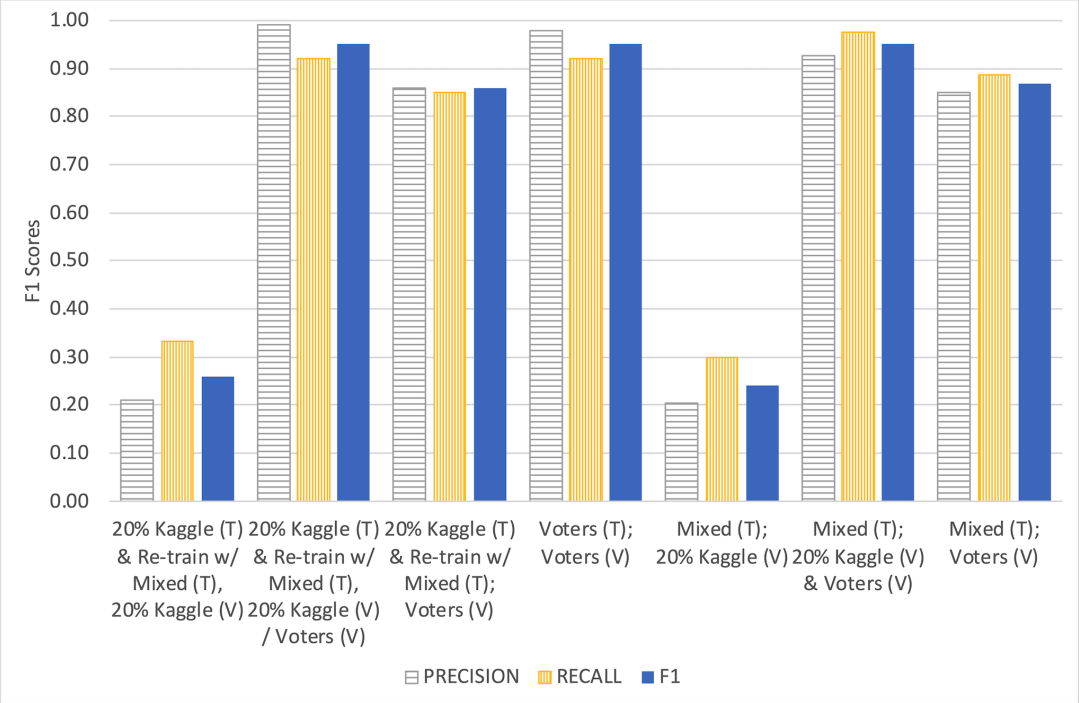


Figure 4.6: $F_1 scores$, Precision and Recall values of the training and re-training sessions with spaCy.

Figure 4.6 shows the scores obtained while evaluating models using datasets from different domains, as well as re-training existent models. In the first case, the model was assessed with the validation section of the 20% Kaggle dataset and the results were disappointing. On the other hand, the second and third cases, using a validation dataset that resembles more to the re-training data, show better results. Therefore, the results of the re-trained the models (first three cases) suggest that models tend to forget previous information and retain more recent data. The fourth case (i.e., in the middle, with voters data) obtained the expected results with over 0.90 $F_1 score$. The last three cases indicate that, in this case, the generalisation capabilities of spaCy models are low. This is visible

in the results of the fifth case, training with mixed data and validating with a section of the 20% Kaggle dataset.

4.3.4 Discussion

Regarding the model-training time with context-specific data, the results have shown that it is not necessary to spend a significant amount of time (or training iterations) to devise a system that is able to correctly identify the intended entities. The longest training session lasted approximately 6500 seconds in spaCy, while Stanford CoreNLP took approximately 1125 seconds and perform similarly in terms of $F_1 score$.

The less positive aspect is the time necessary for manually hand-labelling the data fed to the models. Each document takes an average of 4.75 hours for annotation. The measure indicates the time spent by a person labelling the entities in the referred datasets. Afterwards, each document was reviewed by at least one other person for consistency purposes. Spending approximately 20 hours of manual labelling, already allows the training of a model able to identify entities such as person, city, title, employment details, and others.

It became evident that models perform similarly regardless of having generic or context-specific data. By knowing the behaviour of the machine learning algorithms behind such systems, this is the expected outcome.

Overall, there were no significant differences between Stanford CoreNLP and spaCy classification scores. The main differences are implementation language, model training times and underlying classification mechanism (i.e., CRFs and Neural Networks). Therefore, the candidate tools that are used in the proposed hybrid classification mechanism are chosen based on implementation requirements and overall integration details (described in Section 4.5).

4.4 Model-based Classification

This section analyses the findings of the NER classification approach with ML algorithms. The two algorithms applied in this mechanism were MLP and RF. The choice of algorithms was based on their underlying characteristics (i.e., a Neural Network vs an Ensemble of Decision Trees). There were several other options to choose from (e.g., C4.5, AdaBoost, and others). However, the scope of this analysis was not to make an exhaustive comparison between ML algorithms, but rather choosing at least two that are well tested and documented. As explained before (Section 4.1), word vectors need to be used to convert the text into numerical inputs. This process is a necessary step since the algorithms (MLP and RF) process data in its numerical representation.

Table 4.2 shows the aggregate results of the model training with MLP and RF, using different combinations of datasets and word vectors. In a homogeneous experiment (Group 1), the $F_1 score$s were among the highest (except the results of the word vector pre-trained with the validation file). However, when both

Table 4.2: MLP and RF model training results.

Group	Datasets	Pre-trained Word Vector	MLP			RF		
			Precision	Recall	F_1score	Precision	Recall	F_1score
1	Train: Kaggle (20%) Validation: Kaggle (20%)	Training File	0.87	0.55	0.65	0.85	0.79	0.82
		Validation File	0.03	0.00	0.00	0.03	0.10	0.04
		Google	0.82	0.80	0.81	0.84	0.74	0.79
		Glove	0.80	0.71	0.74	0.80	0.67	0.72
2	Train: Kaggle (20%) Validation: Voters	Training File	1.00	0.11	0.20	0.84	0.89	0.86
		Google	0.54	0.88	0.59	0.45	0.26	0.33
		Glove	0.29	0.58	0.38	0.13	0.24	0.17
3	Train: Kaggle (20%) Val.: Kaggle (20%) and voters	Training File	0.90	0.56	0.67	0.78	0.85	0.81
		Google	0.58	0.85	0.68	0.82	0.62	0.70
		Glove	0.40	0.69	0.49	0.57	0.57	0.55
4	Train: Combined Validation: Voters	Training File	0.20	0.09	0.12	0.15	0.09	0.11
		Google	0.48	0.48	0.48	0.52	0.46	0.49
		Glove	0.53	0.54	0.53	0.53	0.51	0.52
5	Train: Combined Val.: Kaggle (20%) and voters	Training File	0.00	0.01	0.00	0.15	0.01	0.01
		Google	0.24	0.26	0.24	0.25	0.25	0.25
		Glove	0.27	0.29	0.28	0.26	0.28	0.27
6	Train: Combined Validation: Kaggle (20%)	Training File	0.20	0.00	0.00	0.02	0.00	0.00
		Google	0.60	0.40	0.50	0.05	0.03	0.04
		Glove	0.60	0.40	0.50	0.05	0.03	0.03

datasets were combined (Group 5) but from the same pre-trained word vectors, the results were between 0 and 0.3.

Training with datasets from different sources (Group 4 and 6) show that MLP performs better than RF using the pre-trained Glove word vectors. When the models are trained with Kaggle data (Group 2 and 3), RF provided better results.

MLP and RF performances were similar. RF 's highest score was with Kaggle's training data and validation with the voters' dataset (Group 2, first row). MLP 's highest score was with Kaggle dataset (for training and validation) with Google's pre-trained word vector (Group 1, third row).

Since MLP is a neural network, when the classifier has never seen the input before, it will still generate an approximate output. Contrary to the Random Forest classifier, which cannot provide approximation since it is a combination of multiple decision trees and relies on finding the matching node.

Table 4.3: Highest F_1score per algorithm.

Algorithm	Own Word Vector	Google	Glove
MLP	0.67	0.81	0.74
RF	0.82	0.79	0.72

Overall, when trained with general-purpose data (i.e., Kaggle dataset), both classifiers performed well. This also relates to a more extensive vocabulary that the models can recognise during validation. As shown in Table 4.3, the highest score (0.82) was obtained while pre-training word vectors with Kaggle's dataset and classifying with RF. However, the score for MLP is 0.67. Running MLP and RF with Google's pre-trained word vector showed the most promising results: 0.81 and 0.79, respectively.

The above results suggest that pairing Google's pre-trained word vector with MLP and RF is the most appropriate choice. Therefore, the design of the hybrid classification mechanism proposed in the next section takes this into account.

4.5 A Hybrid Classification Approach

The lessons learned from the approaches described in the previous sections motivated the proposal of a novel classification mechanism with the potential to highly benefit data analysis tasks.

The proposed hybrid classification approach is a combined mechanism composed of different data analysis methods. NLP tools, custom-made NER models with MLP and RF, as well as tailored regular expressions, are the core of the approach, as described next.

4.5.1 Architecture

As suggested before, the proposed solution is a combination of classifiers. The objective is to enhance the predicted output by combining the capabilities of individual models. More specifically, the approach considers NLP tools, NER models such as MLP and RF, and regular expressions - thus enabling the detection and classification of the following PII types:

- Title
- Last name
- First name
- Street name
- Street number
- Post code
- City
- Country
- Gender
- Date of birth
- Salary information
- Email address
- Social security number
- Bank details
- License plate number
- National id number
- Passport number
- Phone number
- Credit card number

English is the main language supported by the solution. Nevertheless, additional language models (provided by spaCy) were also deployed. They enable detection of entities in languages such as Spanish, French, and Italian – the languages also used in the four PoSeID-on Use Cases, where the solution is evaluated. Figure 4.7 presents the architecture of the proposed classification mechanism.

The choice of NLP tools was based on the experimental work described in the previous sections. NLTK and spaCy are integrated into the hybrid mechanism.

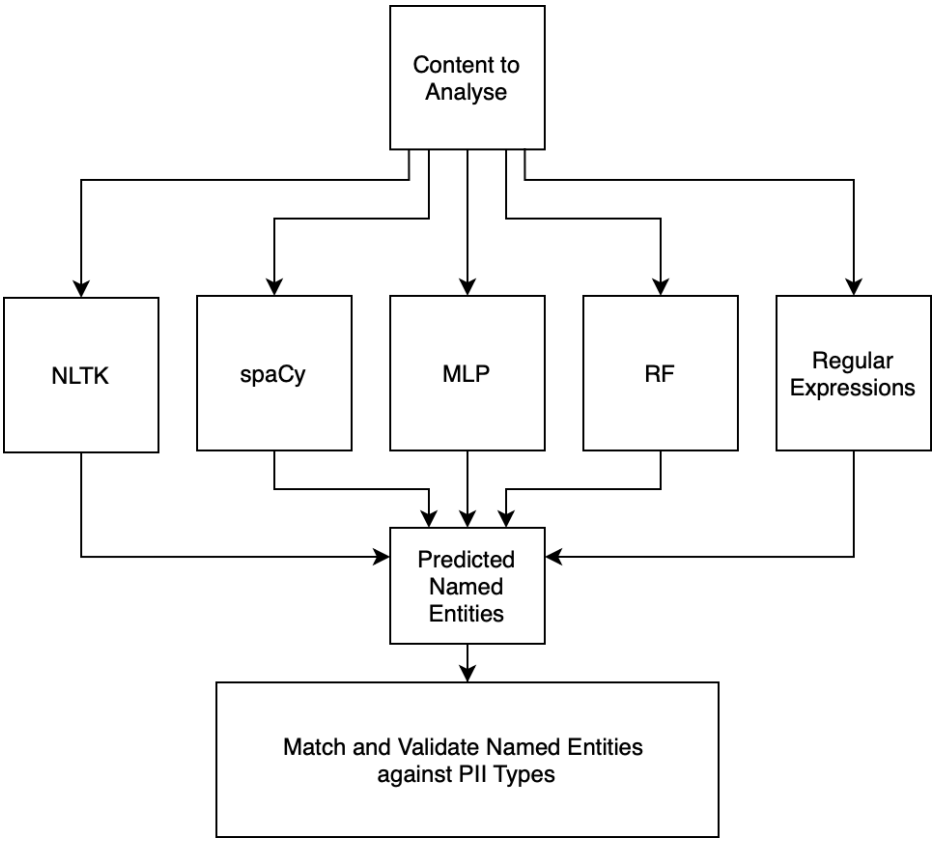


Figure 4.7: Hybrid classification solution.

Although NLTK classification did not perform as well as spaCy or Stanford CoreNLP, it is lightweight and fast, which is also relevant for implementation and deployment purposes. SpaCy classification performed similarly to Stanford CoreNLP. It is developed in Python, and its dependencies are lightweight as well. Stanford CoreNLP, despite being very suitable classification model, is implemented in Java and requires more extensive dependencies, which increases integration complexity with minimal classification gains. Therefore, since the classification performance is similar to spaCy, Stanford CoreNLP was not included in the solution.

MLP and RF algorithms were used to create ML models capable of performing Named Entity Recognition over the previously described PII types. Word Vectors were also necessary to convert text to the numerical representation required by MLP and RF algorithms. Specifically, Google's pre-trained word vector was adopted.

Moreover, Regular Expressions were designed and tailored for the detection of specific PII types that are usually language neutral. For instance, email address, phone number, credit card number or International Bank Account Number (IBAN) are some of those PII types. This approach not only allows analysis over PII types from different languages but also to validate its contents. For instance, validating whether the IBAN is actually valid and not just rely on checking its format.

4.5.2 Validation

The output of the NLP tools, ML models and regular expressions is the prediction of the named entity associated with the input data. In cases where the predicted entities are different for the same data (e.g., NLTK predicts "Person" and MLP predicts "Location"), a voting system is used to choose the final prediction.

The voting system can be set up in several forms. The proposed mechanism currently relies on a majority voting system. Nevertheless, it can be replaced by another voting system, like those considering the models' $F_1 scores$ or a density-based weighting that considers different training data.

Table 4.4: Classification comparison (Accuracy).

PII Type	NLTK	spaCy	MLP	RF	Proposed Solution
Title	0	0	0.54	0	0.54
First Name	0.15	0.34	0.87	0.39	**0.94**
Last Name	0.26	0.2	0.93	0.51	**0.95**
Street Name	0.68	0.03	0.13	0	**0.92**
Street Number	0	0.68	0	0.01	**1**
Post Code	0	0.48	0	0	**0.68**
City	0.71	0.07	0.18	0.03	**0.82**
Country	0.84	0.85	0.76	0.81	**1**
Gender	0	0	0	0	**1**
Date of Birth	0	1	0	0	**1**
Salary Information	0	0.19	0	0	0.19
Email Address	0	0	0	0	**1**
Soc. Sec. Number	0	0	0	0	**1**
Bank Details	0	0	0	0	**1**
Lic. Plate Number	0	0	0	0	**1**
Nat. ID Number	0	0	0	0	**0.89**
Passport Number	0	0	0	0	**1**
Phone Number	0	0.86	0	0	0.86
Credit Card Number	0	0	0	0	**1**
Average (for all types)	0.13	0.24	0.17	0.09	0.90
Average (for supported)	0.52	0.47	0.56	0.35	

Table 4.4 provides a comparison of the results obtained while testing different mechanisms against the proposed hybrid classification approach. The testing set is based on the English language inputs, with a sample of at least 200 records for each PII type. The sample contained 19 different PII types used as in input for the hybrid classifier proposed in this section.

As Table 4.4 shows, the classification accuracy of individual mechanisms is mostly inferior when compared to the proposed mechanism. The highlighted scores, in the last column, achieved higher accuracy scores. In cases such as Title, Salary Information and Phone Number, the scores were equal. Since the

analysis was performed case by case, with binary classification, accuracy is the most relevant metric as it indicates the ratio of correctly classified entities to the total observations. The overall score is 0.90, which suggests an improvement of approximately 60% against the second-highest score (MLP, 0.56). The cases where the proposed solution achieves an accuracy score of one (1) are associated with the employment with the regular expressions.

Additional evaluation results are presented in Chapter 5, where the evaluation under the PoSeID-on Use Cases is discussed.

Next section presents the lessons learned throughout the experimental work described until now.

4.6 Lessons Learned

All actions taken towards the definition, implementation and validation of the proposed mechanism involved a series of steps that allowed conclusions to be derived from the process. Conclusions not only regard NLP, NER or ML but also how the combination of such technologies can be applied to identify PII for the ultimate benefit of the user. The main lessons and findings of this work are highlighted in the next subsections, while the evaluation in the PoSeID-on Use Cases is presented in the next Chapter.

4.6.1 Dataset Size and Classification Accuracy

There were a total of 47,959 sentences (1,354,149 tokens) readily available and labelled, which is a considerable sample size. Dividing the dataset in smaller chunks allowed the assessment of dataset size influence on ($F_1 score$).

Approximately 20% of the total dataset size was sufficient to provide results very similar to the ones obtained using the full-sized dataset. The proportion of 20% is equivalent to 9,590 total sentences, and the $F_1 score$ variation is 0.01, 0.03, and 0.05 for NLTK, Stanford CoreNLP, and spaCy, respectively.

4.6.2 Data Diversification

As previously stated, it is hard to retrieve quality data that matches the outlined requirements. It is only natural since this is precisely the type of information that should be kept out of the public domain.

It was possible to determine that the models' generalisation capabilities are not optimal. Although MLP is able to perform approximations, RF cannot. Nevertheless, the solution proposed in the previous section shows that it can partially overcome this limitation.

4.6.3 Manual-labelling Effort

Manually labelling data is a time-consuming task. Unless there are other sources of already labelled data, it is a necessary task. Nevertheless, since many PII

types are sensitive by default, such datasets (even unlabelled) are not as generally available as other types of data.

It is not feasible to systematically (manually) label large amounts of data. To counter this limitation, one may consider two possibilities. The first is to assess the feasibility and reliability of using a synthetic data generator (e.g., Mostly AI [MostlyAI, 2020]). The second possibility is to recur to online annotation services [Neves and Ševa, 2019]. However, the latter still depends on finding appropriate and sufficient datasets, which is equally challenging.

Nevertheless, future work direction is likely to consider further analysis of the aforementioned options.

4.6.4 Data Validation and Discovery

Data validation is applied to a variety of services and fields. From verification of text boxes in web pages to more complex processes that ensure the delivery of clean and valid data. With the proposed approach, systems should be able to validate not only data types and formats but also contents, in a privacy-preserving manner.

Systems managing text data inputs (such as forms that collect some form of PII) are able to distinguish if the input matches the actual description and whether the content is valid. For instance, systems would be able to generate a warning if IBAN or social security number are not correct. Additionally, validating open text fields whenever they are filled with sensitive data would also be possible, thus avoiding the submission of sensitive information in undesirable circumstances. Some of these verifications can be performed on the level of the user interface, field by field. However, there are scenarios, like bulk data analysis, where the proposed mechanism has the upper hand.

The discovery of PII is closely linked to the scenario described above, as it not only allows data validation but the discovery of previously unidentified PII. This kind of monitoring can be applied in several scenarios, such as transactions or information exchanges between systems and users, documents or databases. As most data processing tasks are associated with sensitive data, this approach allows the system to warn users (when they are directly involved), so they could take appropriate actions. The real-world implementation of such possibility is described in Section 5.3.

4.7 Summary

This chapter proposed a hybrid classification solution that allows the automation of data analysis while maintaining the characteristics of a privacy-preserving system. Such solution, as described in Section 4.5, consists of an ensemble classifier that includes AI-based approaches such as NLP tools (Section 4.3) and other ML mechanisms (Section 4.4). This solution has been integrated in the PoSeID-on H2020 Project platform and evaluated in the scope of that project's Use Cases, as described in the next Chapter (Section 5.3). Furthermore, there

was a discussion of the lessons learned and how the proposed approach can effectively act as a Privacy Enhancing Technology.

The outcomes of this chapter resulted in the following publications:

- Silva, P. et al. (2020). **Risk Management and Privacy Violation Detection in the PoSeID-on Data Privacy Platform**. SN COMPUT. SCI. 1, 188. DOI: 10.1007/s42979-020-00198-9.
- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2020). **Using NLP and Machine Learning to Detect Data Privacy Violations**. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, pp. 972-977. DOI: 10.1109/INFOCOMWKSHPS50562.2020.9162683.
- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2020). **Using Natural Language Processing to Detect Privacy Violations in Online Contracts**. In Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC '20). Association for Computing Machinery, New York, NY, USA, 1305–1307. DOI: 10.1145/3341105.3375774.
- Casaleiro, R. et al. (2020). **Protection and Control of Personally Identifiable Information: The PoSeID-on Approach**. Journal of Data Protection & Privacy, 3(2).

The following chapter presents privacy risk-assessment mechanisms that increase privacy preservation. That outcome is achieved by performing a privacy risk analysis of data transactions in Cloud environments from characteristics such as PII sensitiveness, correlation, Data Processor reputation and others.

Chapter 5

Privacy Risk Assessment Mechanisms

Contents

Privacy risk can be assessed from a variety of perspectives. Nevertheless, common elements typically include sensitiveness of data, correlation between data types, retention times or data processor reputation.

Although many other aspects may influence privacy risk, defining a model capable of inferring privacy risks based on these common elements can significantly benefit data owners and data subjects. This chapter proposes a multi-input privacy risk model and shows how it was validated on real-world Use Cases under the PoSeID-on H2020 Project.

The outcomes of the chapter resulted in the following paper:

- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2021). **Privacy Risk Assessment and Privacy-preserving Data Monitoring**", IEEE Access, 2021 (submitted, under review).

5.1 Introduction

Assessing privacy risks is made possible by analysing a combination of factors or circumstances around specific events. The most common elements in most data transactions are the type of data, size, and duration. Analysing these elements grants the possibility of estimating the likelihood of privacy risks ever happening. The mechanisms proposed in this chapter not only analyse the referred factors but also consider the analysis of involved parties and the sensitiveness of data types involved, as well as the correlation between data types with a custom privacy risk matrix.

The analysis of the elements mentioned above can be performed in a variety of ways. In this chapter, two different methods are proposed. The first is through a crisp model and the second through a fuzzy model. These two mechanisms allow two different types of analysis. Figure 5.1 shows the difference between the outputs of crisp and fuzzy models. Considering the problem of inferring the existence of privacy risk, the crisp model provides a binary answer without showing quantitative analysis. The fuzzy model, on the other hand, quantifies the amount of privacy risk, normalised between 0 and 1.

The internal mechanisms of the proposed fuzzy models are based on membership functions that assess to what degree the real number (i.e., input) satisfies the desired property (e.g., output: extremely high risk or moderate risk, as shown in Figure 5.1). One of the most common functions is the triangular membership function [Sadollah, 2018]. A discussion and a theoretical explanation on why triangular (as well as trapezoidal) membership functions work so well are presented in [Barua et al., 2013]. For these reasons, as well as their fast computation time, these two functions are used in the proposed privacy risk assessment mechanism.

The related work, open issues and contributions are described next. The remaining of the chapter presents the proposed mechanisms (Section 5.2) and shows

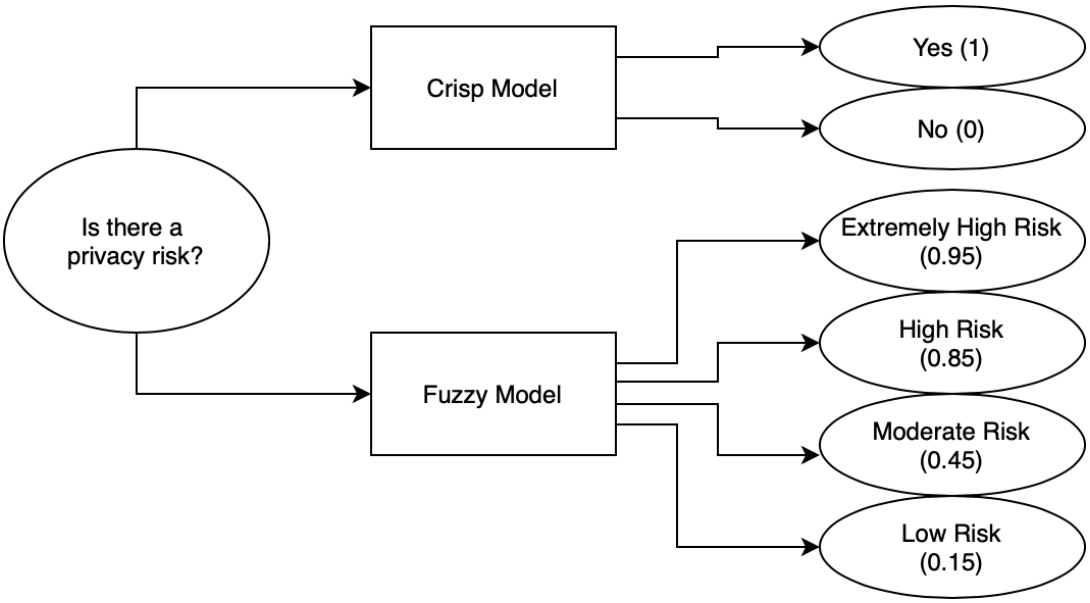


Figure 5.1: Difference between crisp and fuzzy models.

how they were validated in a real-world environment through integration with the PoSeID-on's Cloud platform (Section 5.3).

5.1.1 Related Work and Open Issues

Systems and methods for privacy risk analysis are an essential contribution towards minimising or avoiding privacy breaches in Cloud platforms. Literature [Tang et al., 2016] shows that different mechanisms and approaches offering privacy risk analysis have been proposed over time. Some methods measure privacy risk based on the number of records stored in a system [Grosso et al., 2014] or even based on characteristics of the system [Todd et al., 2016].

Some methods consider expert opinions for the privacy risk assessment. A method proposed by ENISA [ENISA, 2009] relies on expert opinions for privacy risk assessment. Similarly, a semi-quantitative risk assessment framework [Saripalli and Walters, 2010] considers experts opinions and uses statistical data of a system to predict the impact and probability of a privacy risk event.

Other approaches [Dhasarathan et al., 2015; Wang et al., 2012] rely on entire frameworks for privacy risk assessment. However, such kind of approaches can negatively impact the performance of Cloud systems, leading to a deterioration of user experience and service quality because a privacy-preserving scheme usually affects the performance of Cloud systems. As privacy is not free, mechanisms that minimise implementation costs while providing privacy assurances are very challenging.

Adding to the challenge of assessing privacy risks on Cloud systems are the characteristics associated with Cloud services: (1) data growth is exponential; (2) high costs of providing data privacy; (3) the vulnerabilities often found in Cloud systems that lead to data breaches. Moreover, human error, present in

all situations, highly increases the complexity of guaranteeing privacy-preserving Cloud systems. These types of issues are particularly difficult to address. Nevertheless, the contributions presented in this Chapter (described next) reduce, and potentially avoid, privacy risks.

5.1.2 Contributions

Most Cloud systems or services strive to provide secure and private infrastructures. They offer state-of-the-art security mechanisms and (in Europe, with the contribution of GDPR) they are also adapting their privacy policies. Nevertheless, they usually fail in one specific aspect: directly engaging with users for privacy-related aspects.

The mechanisms presented in this chapter allow a direct engagement with users by offering a quantification of privacy risk levels associated with data transactions. This is achieved by a multi-input mechanism that considers the data types involved, data processor reputation, data sensitiveness and correlation, and retention time. Moreover, the proposed methods are validated in real-world scenarios through the integration and deployment in PoSeID-on's Cloud platform.

5.2 Multi-input Privacy Risk Assessment Mechanisms

The following sections describe the primary inputs of the proposed privacy risk assessment mechanisms: reputation, data sensitivity, correlation, number of PII types, retention time and entity matching information. The latter relates to the mechanisms proposed in the previous chapter (i.e., data analysis mechanism for PII identification).

5.2.1 Reputation Assessment

Reputation assessment is the process of collecting, aggregating and distributing data about an entity. Data aggregation is based on literature specification [Vavilis et al., 2014] of requirements and features of reputation systems. Such an assessment can later be applied to predict future behaviours. As there is a constant evaluation of the entities' behaviour, positive ratings are reflected by good performance history and expected system behaviours. In contrast, negative ratings are associated with problems, errors, malfunctioning, etc. Thus, reputation systems not only encourage good behaviour from entities, as also provide users with valuable information in regards to whom or what they will trust their data.

The proposed privacy risk assessment mechanisms can rely on any reputation system output as long as it is normalised. Therefore, the input for reputation ranges from 0 (lowest reputation) to 1 (highest reputation). Figure 5.2 shows how the model assumes four distinct levels of reputation: poor [0 - 0.4], average]0.4 - 0.6], good]0.6 - 0.8] and excellent]0.8 - 1]. Higher reputation entities are deemed as more trustworthy and less likely to raise privacy concerns for users.

On the other hand, lower reputation entities may be more likely to raise privacy concerns.

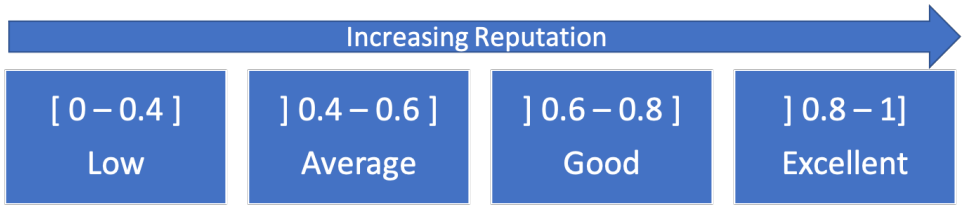


Figure 5.2: Reputation levels.

As the next section shows, reputation is not the only metric the proposed models uses to assess the privacy risk. Nevertheless, it is essential to keep the reputation information as up to date as possible, as it highly affects the privacy risk estimation.

5.2.2 PII Sensitivity and Correlation

PII sensitiveness measures the sensitivity of a PII type [Milne et al., 2017]. It is an essential element on the privacy risk analysis. The sensitiveness level is verified according to the degree of unique identification that each type holds about a data subject. Figure 5.3 proposes a five-level risk classification hierarchy.

PII Risk Description	Risk Level
A - Sensitive (direct identifier)	**HIGH**
B-O - Almost sensitive quasi-identifier (almost direct identifier)	**MIDDLE-OVER**
B - Quasi-identifier (indirect identifier)	**MIDDLE**
B-L - Far quasi-identifier (indirect identifier)	**MIDDLE-LOW**
C - Limited effect	**LOW**

Figure 5.3: PII sensitivity levels.

There is a vast number of PII types. Nevertheless, there is a set of types that are most commonly used in a variety of scenarios. Table 5.1 shows the PII types defined for validating the mechanism under real-world Use Cases from the PoSeID-on H2020 Project, as described in Section 5.3.

It is possible to observe in Table 5.1 that approximately 50% have low risk associated (e.g., title, first name or street name). This low risk is due to the generic nature of those PII types, which can be associated with many data subjects. However, other PII types can uniquely identify data subjects, and in this case, these types have a higher risk level (e.g., Social Security Number or National ID Number).

Table 5.1: Sensitivity level per PII type.

PII Type	**Risk Level**	**PII Type**	**Risk Level**
Title	C	Social Security Number	A
First Name	C	Bank details	A
Last Name	B-L	Employment Contract & Salary Info.	A
Email Address	A	License Plate Number	A
Street Name	C	Gender	B-L
Street Number	C	National ID Number	A
Post Code	C	Date of Birth	C
City	C	Phone Number	B-L
Country	C	Passport Number	A

Although Table 5.1 associates risk levels to individual PII types, it is also possible to have combined and cumulative risk levels. There are cases where several types are analysed or exchanged. In those cases, it is necessary to assess the sensitiveness levels of combinations of PII types. Figure 5.4 shows risk levels of 2-by-2 combinations. All PII types with maximum risk level (i.e., "A") are excluded from the table as any combination would generate the maximum risk level for every pair. The reasoning behind the combined risk assessment is that each pair always inherits the highest level from its elements (as the minimum aggregated risk level of the combination).

Inevitably, there are cases when more than two PII types are analysed. The verification of the progressive and cumulative risk levels is shown in Figure 5.5. This matrix shows a progressive and cumulative risk association among PII types. Pivoting from left to right allows assessing which cumulative set of data eventually results in the highest risk level. Since such a risk level can only be achieved when data are matched with "First Name" and "Last Name", all other types need not be considered, at this point.

The following section describes two other inputs: retention time and data analysis. The latter considers the data analysis mechanisms discussed in Chapter 4.

5.2.3 Retention Time and PII Analysis

There are two types of retention times: A - the period which data is kept for the service to be provided; B - the period which data is kept for legal purposes after the service is provided.

GDPR does not specify retention times for the processing of personal information while services are being provided. Nevertheless, it states that personal data may only be kept in such a way that allows identification of citizens for no longer than is necessary for the purposes for which it was processed. This refers to case one (A), described above. The second case (B) refers to the amount of time an

PII Fields	Risk Level	Title	Last name	First name	Street name	Street number	Post code	City	Country	Gender	Date of Birth	Phone number
Title	C		B-L	C	C	C	C	C	C	B-L	C	B-L
Last Name	B-L	B-L		B	B-L	B-L	B-L	B-L	B-L	B-L	B	B-O
First name	C	C	B		C	C	C	C	C	B-L	C	B
Street name	C	C	B-L	C		C	C	C	C	B-L	C	B-L
Street number	C	C	B-L	C	C		C	C	C	B-L	C	B-L
Post code	C	C	B-L	C	C	C		C	C	B-L	C	B-L
City	C	C	B-L	C	C	C	C		C	B-L	C	B-L
Country	C	C	B-L	C	C	C	C	C		B-L	C	B-L
Gender	B-L	B-L	B-L	B-L	B-L	B-L	B-L	B-L	B-L		B-L	B-O
Date of Birth	C	C	B	C	C	C	C	C	C	C		B-L
Phone number	B-L	B-L	B-O	B	B-L	B-L	B-L	B-L	B-L	B-O	B-L	

Figure 5.4: PII risk correlation matrix.

PII Fields	Risk Level	Title	Last name	First name	Street name	Street number	Post code	City	Country	Gender	Date of Birth	Phone number
Last Name	-	B-L		B	B	B	B	B	B	B-O	A	B-O
First name	-	C	B		B	B	B	B	B	B-O	A	B-O

Figure 5.5: PII cumulative risk levels.

entity is requested, by law, to keep citizens data.

Since case one (A) varies according to the length of the service being provided, it is not feasible to determine whether an amount of time is excessive or not. Nevertheless, based on the legal framework by which entities should retain data after the service is provided, it is possible to verify if the initial retention request is in line with local regulations or not. For instance, in Germany, post-service retention times vary between four weeks and three years. As such, when an entity is requesting to keep data for over three years, this can be considered abuse. Therefore, the proposed multi-input risk assessment mechanism considers this perspective of retention times in its privacy risk evaluation. Table 5.2 shows the retention times defined in the countries of some of the partners involved in the PoSeID H2020 Project (described in Section 5.3).

Another aspect considered by the mechanism is PII analysis. It is assumed that a previous analysis is performed on the data, for instance, with the mechanisms proposed in the last chapter, stating whether exchanged data was validated or not. In cases where such analysis is not available, the contents are expected to be valid.

Table 5.2: Retention times of some European countries (involved in the PoSeID-on Project).

Country	**Retention Times**
Germany	1 month to 3 years
Italy	5 to 10 years
Spain	1 month to 30 years
France	1 month to 10 years
Portugal	1 month to 20 years
Malta	5 to 10 years

The next sections describe how the proposed mechanism handles the inputs mentioned above in two different ways: crisp and fuzzy approaches. The former provides static and categorical privacy risk assessment. In contrast, the latter supports linguistic conversion and provides a quantification of the privacy risk level.

5.2.4 Crisp Model

The objective of this model is to assess the likelihood of a systems' privacy-breaching event. The output can be positive (i.e., high privacy risk) or negative (i.e., low privacy risk). Crisp sets are an effective approach to model such systems, as they are based on binary logic. Therefore, crisp decision making is the basis of this model.

As mentioned before, several elements influence the decision making. Figure 5.6 shows how these elements influence the model's decision making: retention time (A), reputation (B), PII sensitiveness (C), number of PII (D) and their correlation (E).

The most significant elements are reputation (A) and PII sensitiveness (C). A decrease in reputation (A) increases privacy risks. Similarly, an increase in PII sensitiveness also increases privacy risks.

The proposed privacy risk assessment model indicates, in a binary way, the likelihood of a privacy-breaching event. It is a simple yet effective way of assessing privacy risks. The model does not quantify the probability of those events ever happening. Similarly, it does not guarantee that such events will ever occur. Nevertheless, it is a convenient way of raising privacy awareness for users of systems where the model is deployed.

In cases where quantification of privacy risks is more adequate, a binary model is not sufficient. Accordingly, the next section proposes a fuzzy model that translates categorical linguistic inputs into real numbers, better representing the uncertainty associated with the inputs described in Sections 5.2.1, 5.2.2 and 5.2.3. Moreover, it quantifies and normalises the output in the form of a real number between 0 and 1.

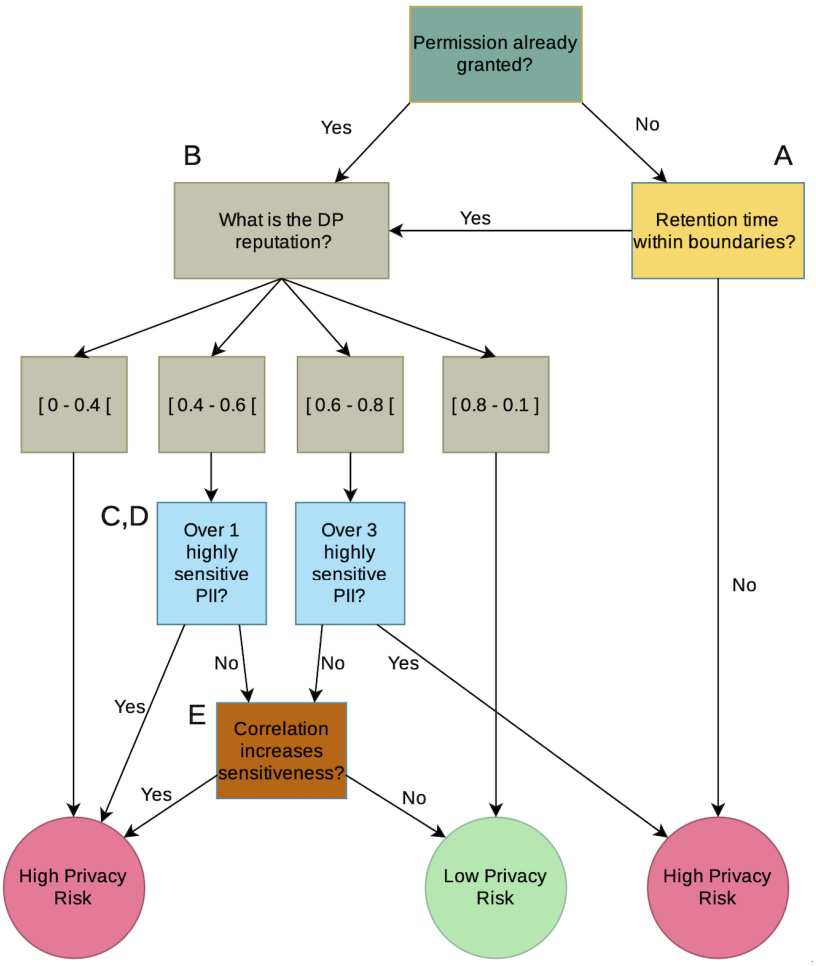


Figure 5.6: Privacy risk assessment model – crisp approach.

5.2.5 Fuzzy Model

As seen before, crisp models are useful, as an initial concept, to model the digital systems working on binary logic. However, they fall short in offering quantification and categorical translation of inputs and outputs. A binary model, like the one proposed before, triggers privacy risks alarms when the probability of such an event is elevated. However, the model is not capable of assessing the level of risk associated with each event. Risks events with 99% or 60% probability have the same output. This limitation can be surpassed with Fuzzy Logic mechanisms.

Fuzzy models intend to introduce imprecision and vagueness not found in binary models. The model's inputs can be translated from crisp (i.e., precise representation like "0.5") into categorical representations (i.e., linguistic representation like "average"). This is the fuzzification process.

Similarly, the output can be converted from categorical into a crisp representation (i.e., any real number between 0 and 1). This is a process designated defuzzification. Privacy risk events generated by a fuzzy model can, therefore, be classified according to their severity level based on their crisp or categorical

representation.

As previously mentioned, literature [Barua et al., 2013; Sadollah, 2018] shows that triangular and trapezoidal membership functions are suitable for a variety of tasks. As such, the proposed mechanism employs those functions.

The inputs of the proposed model and the respective fuzzification process are described next.

Retention Time

The crisp values associated with this input range between [0 - 100], as shown in Section 5.2.3). This allows mapping any range of retention times to a normalised value. The fuzzy representation is described by the following sets: very short, short, average, long, excessive.

Figure 5.7 shows the trapezoidal functions defined for each fuzzy set of retention time. The kernels' edges overlap with each of its neighbours to include the degree of uncertainly provided by fuzzy logic. The function highlighted in red represents the first set (i.e., very short).

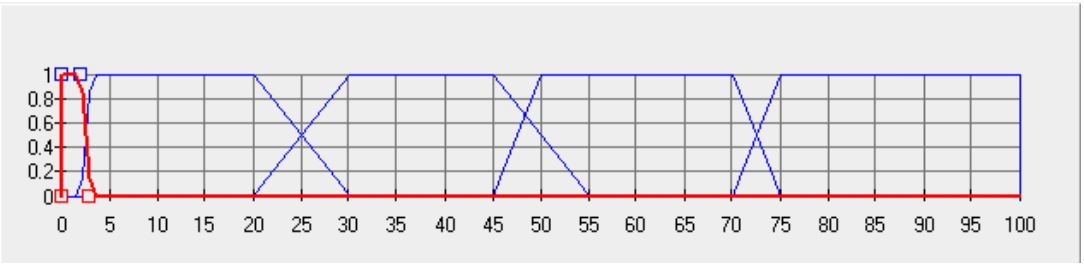


Figure 5.7: Retention time – membership functions.

Data Processor Reputation

The crisp values associated with this input range between [0 - 1], as shown in Section 5.2.1). Similarly to the previous, this allows mapping any kind of reputation system to a normalised value. The fuzzy representation is described by the following sets: poor, average, good, excellent.

Figure 5.8 shows the trapezoidal functions defined for each fuzzy set of reputation. The kernels' edges overlap with each of its neighbours to include the degree of uncertainly provided by fuzzy logic. The function highlighted in red represents the first set (i.e., poor).

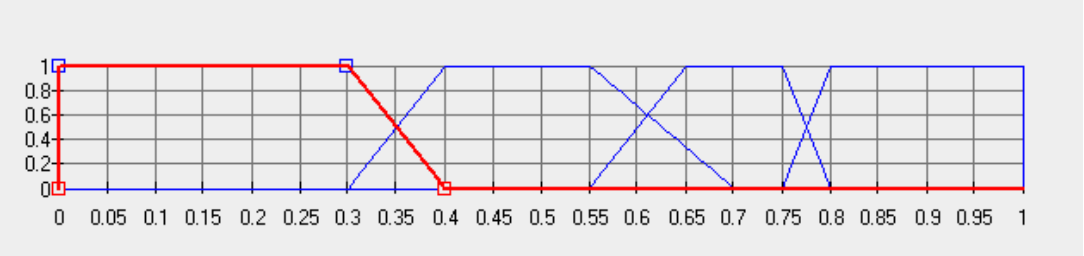


Figure 5.8: Reputation – membership functions.

PII Sensitiveness

The crisp values associated with this input range between [0 - 6]. In this case, the input is associated with the sensitiveness levels as described in Section 5.2.2. The fuzzy representation is described by the following sets: high, middle-over, middle, middle-low, low.

Figure 5.9 shows the triangular functions defined for each fuzzy set of PII sensitiveness. The kernels' edges overlap each of its neighbours, up to their centre, to include the degree of uncertainly provided by fuzzy logic. The function highlighted in red represents the first set (i.e., high).

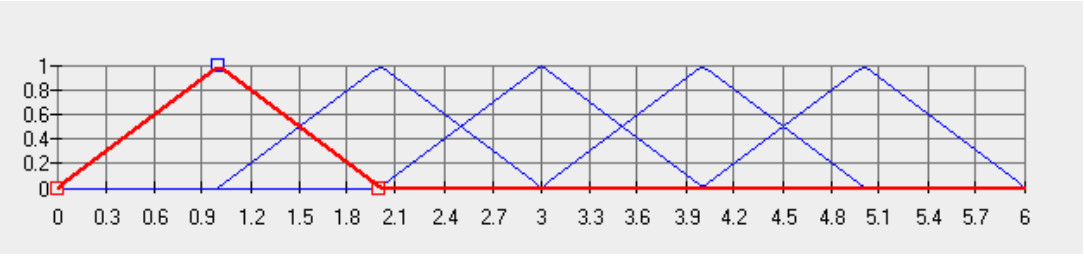


Figure 5.9: PII sensitiveness – membership functions.

PII Correlation

The crisp values associated with this input range between [0 - 6]. Similar to the previous case, here, the input is associated with the correlation levels as shown in Section 5.2.2. The fuzzy representation is described by the following sets: high, middle-over, middle, middle-low, low.

Figure 5.10 shows the triangular functions defined for each fuzzy set of PII correlation. The kernels' edges overlap each of its neighbours, up to their centre, to include the degree of uncertainly provided by fuzzy logic. The function highlighted in red represents the first set (i.e., high).

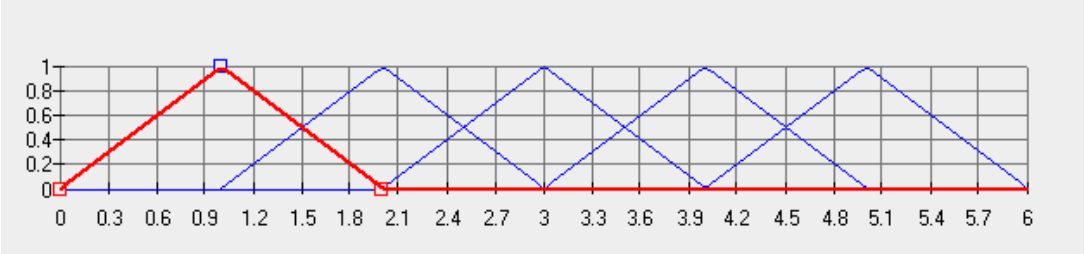


Figure 5.10: PII correlation – membership functions.

Number of PII Types

The crisp values associated with this input range between [0 - 19]. This is the number of PII types previously described in Section 5.2.2. The fuzzy representation is described by the following sets: none, few, middle, high, very-high.

Figure 5.11 shows the trapezoidal functions defined for each fuzzy set of reputation. The kernels' edges overlap with each of its neighbours to include the degree of uncertainly provided by fuzzy logic. The function highlighted in red represents the first set (i.e., none).

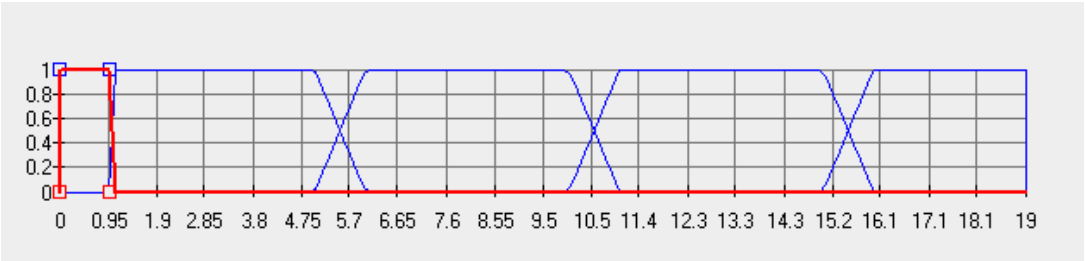


Figure 5.11: Number of PII types – membership functions.

There are at least two reasons for using triangular and trapezoidal membership functions instead of asymmetric, Gaussian or other functions. The first is the fact that they require a small amount of data to be defined. Secondly, the fine-tuning of parameters is not as complex as for other membership functions. These two reasons allow this mechanism to be easily adapted and fine-tuned to different scenarios where the inputs might vary (e.g., number of PII types).

The fuzzy relationship between inputs and output cannot be assessed merely with individual input rules. Therefore, such a relationship is assessed by the aggregation and overlapping of the previously described individual input rules, resulting in a more encompassing and complete coverage. The inference method is *Perception-based Logical Deduction*, and defuzzification is based on *Simple Defuzzification of Linguistic Expressions*. Figure 5.12 shows the model schema.

As shown in Figure 5.12, the linguistic output of the proposed model has 7 levels, from *Extremely High Risk* to *Extremely Low Risk*. Each of these levels is associated with a crisp (i.e., numerical) value. For instance, *Extremely High Risk* can be associated with 0.95.

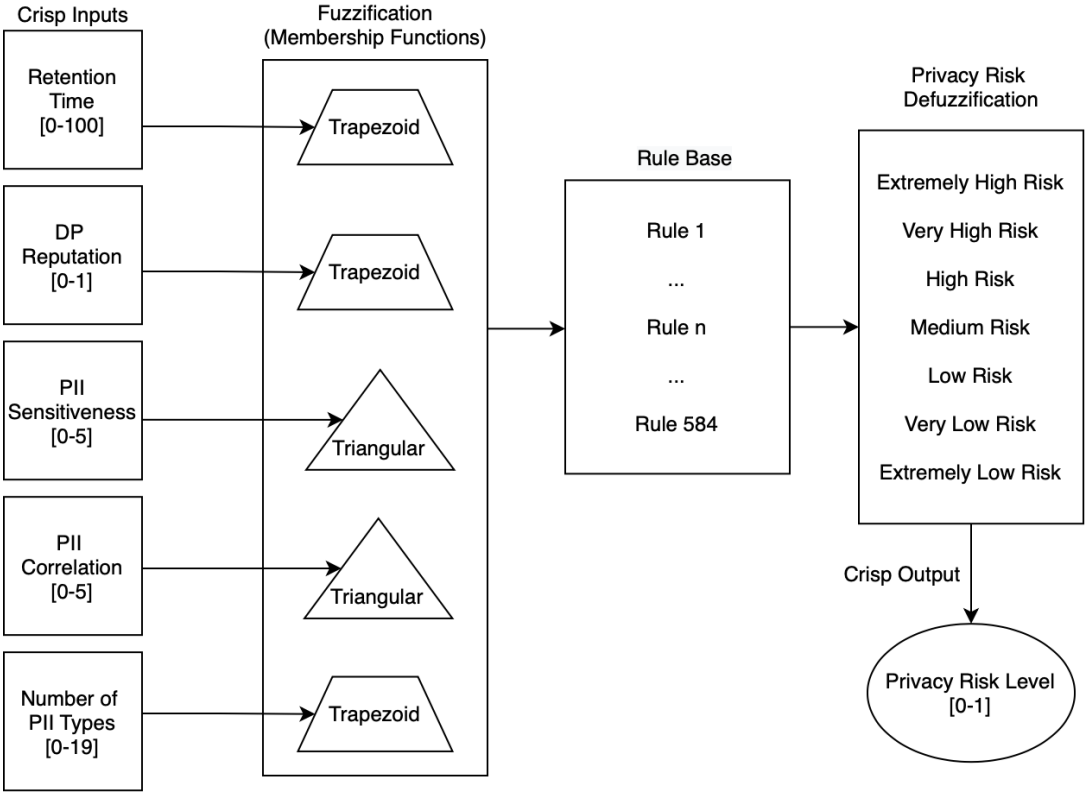


Figure 5.12: Privacy risk assessment model – fuzzy approach.

Another striking feature of the proposed model is that, according to the inputs provided, each categorical output can result in different numerical outcomes. Suppose there are two different sets of inputs that result in *Very High Risk*. One input may become associated with 0.88 and other with 0.86.

This differentiation is made possible due to a comprehensive rule base that assesses 584 individual situations. The rule base considers restrictions such as non-repetition and non-allowed combinations. Each rule considers the linguistic inputs of the model and derives a linguistic output. Table 5.3 shows the fuzzy and crisp privacy assessment associated with specific inputs.

Table 5.3: Linguistic versus crisp input.

Input Type	**Retention**	**Reputation**	**Sensitiveness**	**Correlation**	**Number of PII**	**Output**
Linguistic	Short	Excellent	High	High	Middle	Small Risk
Crisp	25.48	0.82	1.15	1.15	11	0.10

The following section describes how the mechanisms proposed in this and the previous chapters were integrated in the PoSeID-on's Cloud platform.

5.3 Integration and Evaluation in the Scope of the PoSeID-on Cloud platform.

The PoSeID-on project developed an innovative Cloud platform aimed at safeguarding the rights of European citizens. It does so by combining technologies such as Smart Contracts, Blockchain and a web-based dashboard, as well as privacy and risk analysis modules. Overall, it supports organisations in managing and processing data while ensuring compliance with GDPR regulations.

5.3.1 Architecture

Figure 5.13 shows the PoSeID-on architecture. Two of the core modules of the platform are the PDA, for privacy risk analysis, and the Risk Management Module (RMM), for security risk analysis. The mechanisms proposed in Chapter 4 and 5 have been used to build the PDA module.

The PDA is used to monitor personal data transactions, when explicit permission is granted, in order to detect and prevent privacy risks and misbehaved transactions. A warning is issued to the user every time privacy risk thresholds are triggered. There is no data collection, at any moment. Only passive analysis for the benefit of the user, whenever the user chooses to do so. All data is discarded after each analysis. In case a user does not provide explicit consent (which is fully optional, in the scope of the PoSeID-on platform), the PDA does not operate for that specific user. Figure 5.14 shows the architecture of the PDA.

By directly connecting to RabbitMQ (PoSeID-on's shared message queuing service) the PDA establishes communication with two other modules: the Web based Dashboard and the Data Processor API. This allows the PDA to receive analysis requests and analyse whenever there is explicit consent from the data subjects. In addition, a shared REDIS database provides Data Processor (DP) reputation information which is used in the privacy risk assessment mechanism.

5.3.2 Functionalities

The hybrid data analysis mechanisms, presented in Section 4.5, and the multi-input privacy risk mechanism, presented in Section 5.2, support two main scenarios:

- *Privacy Risk Assessment Upon Data Processor Permission Request* – This is associated with the initial interaction between a Data Processor and a Data Subject. Every time a data processor wants to have personal data from a data subject, it needs to formally request permission to use the specific data subject's PII types it wants.

  In this case, the PDA analyses the PII types being requested, the reputation of the requesting Data Processor, and the retention time. Based on

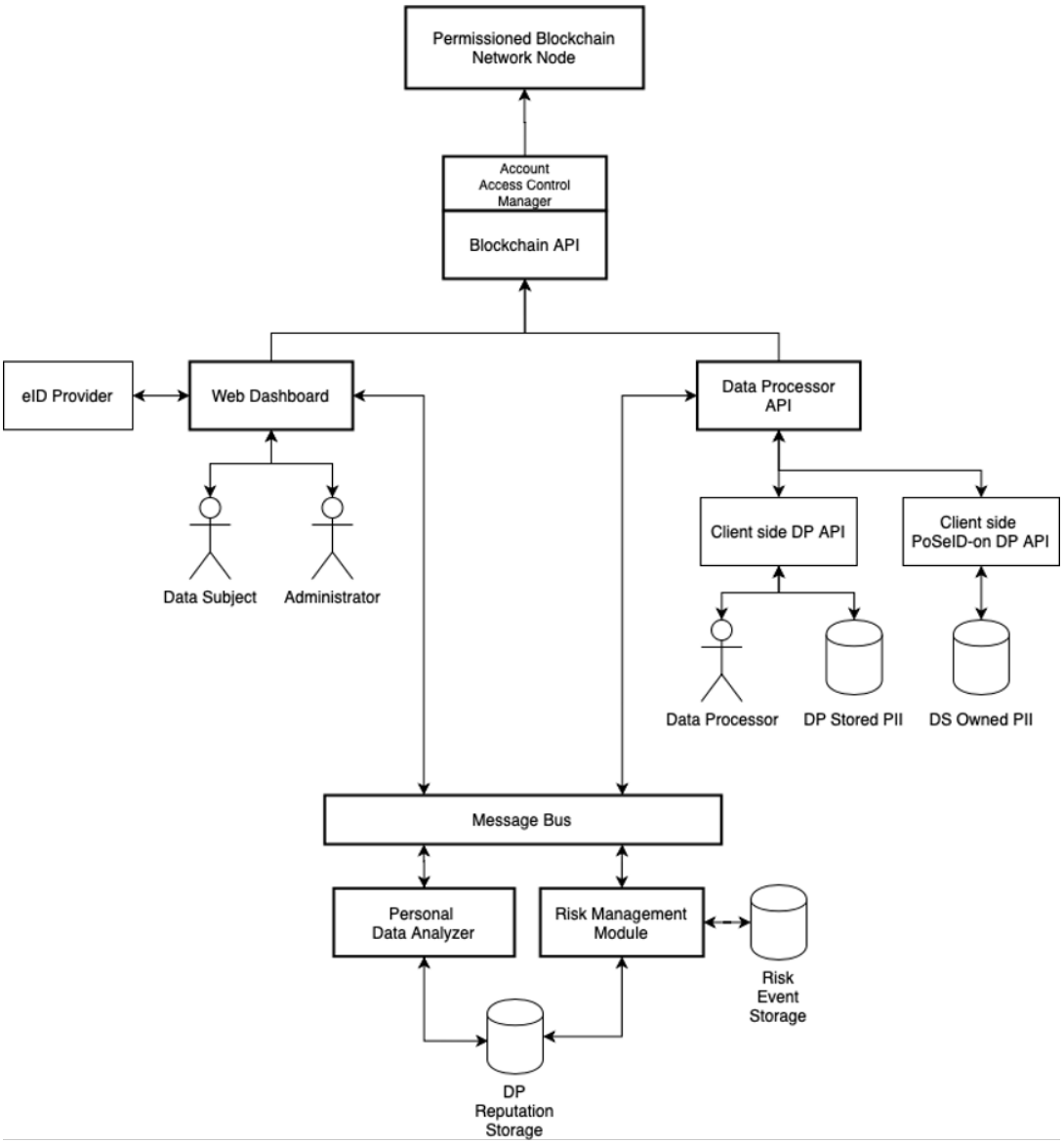


Figure 5.13: PoSeID-on platform architecture.

this analysis, a warning message is sent to the data subject whenever a privacy risk threshold is crossed.

- *Privacy Risk Assessment After Permission is Granted* – After permission has been granted, the PDA can still monitor transactions and assess privacy risks. Typical interactions after granting permission include access to the data or updating its contents.

  In this scenario, the PDA analyses the current reputation of the Data Processor, the data sensitivity and the contents of the PII types involved (with the mechanisms proposed in Chapter 4). Every time data is not dully validated, the reputation decreases or the involved data types are highly sensitive, a warning message is sent to the data subject.

As mentioned above, in case privacy risk thresholds are crossed, warning mes-

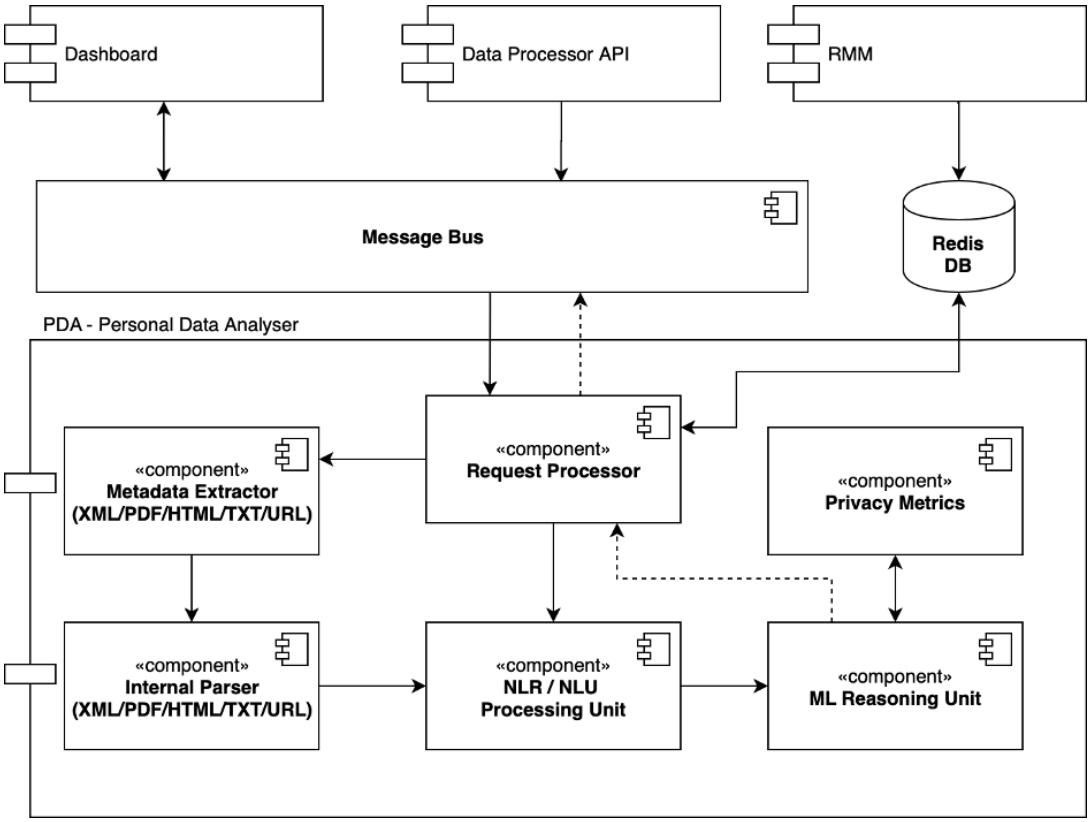


Figure 5.14: Personal Data Analyser (PDA) architecture.

sages are generated and sent directly to the data subject inbox. These messages are sent through the PoSeID-on shared message queuing service and displayed in the Dashboard. As seen in Figure 5.15, the warning message contains the description of the event, the reason of the warning and a button to directly navigate to the Data Processor page, allowing for a quick and direct action.

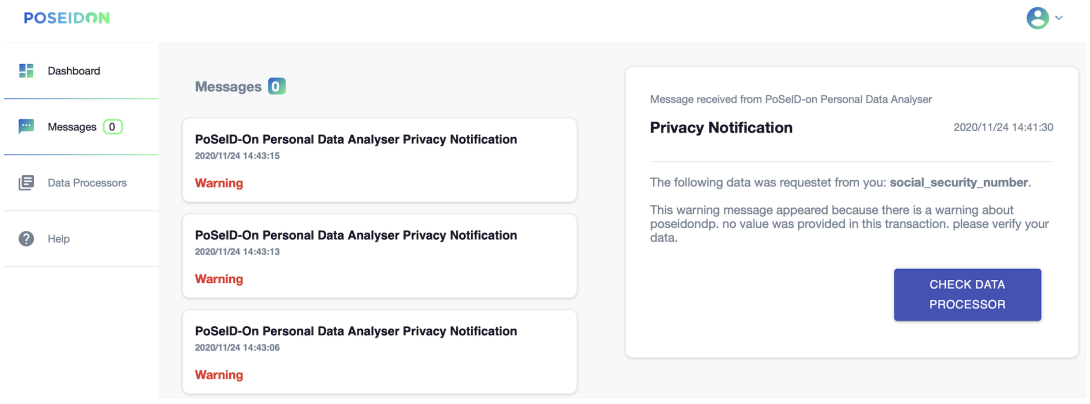


Figure 5.15: Warning message issued upon request permission.

Such warnings enhance the users' chances of avoiding situations where privacy might be compromised. They enable the user to take quick and effective actions such as revoking the permissions previously granted to a specific data processor. Moreover, they raise privacy awareness in measurable and straightforward terms.

The previously described capabilities are translated into a specific set of functionalities in the scope of the PoSeID-on platform. The supported core functionalities are the following:

- *Request Permission* – This function analyses the permission request that a Data Processor sends to a Data Subject. It considers the PII types being asked, the retention period for the permission to be active and the data subject public certificate.
- *Get PII Types* – In this case, the permission is already granted. However, since the Data Processor reputation can change over time, this functionality analyses the data processor reputation and the sensitiveness of the PII types.
- *Get PII* – This function is an effective data exchange, either from a Data Subject or from a Data Processor. While in the first two functionalities the PDA's data analysis capabilities (i.e., mechanisms described in Chapter 4) are idle, in Get PII (and the following two) the module uses all the capabilities of its hybrid classification approach to identify and validate personal information.
- *Update PII* – The update PII functionality is similar to "Get PII". However, in this case the analysis is performed over data that was recently changed and not on data that might have been stored for longer periods.
- *PII Analysis* – The PII analysis is a functionality that allows Data Subjects and Data Processors to submit data for an automated analysis. One of the use cases is when Data Processors send files (e.g., TXT or PDF) for analysis.

5.3.3 Evaluation

Table 5.4 shows the functionalities of the module, their input parameters, and the approximate frequency of usage of each functionality (based on proportions used in the test scripts and on typical usage scenarios of the PoSeID-on platform). The last functionality ("Registration Notification") refers to handling the notification that the platform sends every time a new user registers. Thus, allowing the module to formally request user permission and activate its functionalities as soon as consent is granted.

Table 5.5 indicates the average processing time for each functionality. Additionally, it also considers the throughput per instance, as well as with five instances deployed. The values refer to a setup with 12 GB of RAM, six Central Processing Unit (CPU) and 100 GB Solid State Drive (SSD). As evidenced in Table 5.5, the functionalities where data classification pipeline is not activated are very fast, taking approximately 0.03 seconds per transaction.

On the other hand, when the NLP pipeline is active, each transaction takes an average of 1.13 seconds to be processed. This includes performing NER classification with tools such as NLTK and spaCy, as well as loading NER models

Table 5.4: Module functionalities, message parameters and expected volume.

Functionality	**Parameters**	**Amount**
Request Permission	PII Types, Until, Reason, DS Cert	15%
Get PII Types	PII Types, DS Cert	35%
Get PII (DS to DP)	PII Type, DS Cert, Value	17, 5%
Get PII (DP to DP)	PII Type, DS Cert, DP Name, Value	17, 5%
Update PII	PII Type, Value, DS Cert	8%
PII Analysis (DP)	PII Types, Value, File	2.5%
PII Analysis (DS)	PII Types, Value, File, DS Cert	1.5%
Registration Notification	DS Cert	3%

based on MLP and RF. As such, the performance results have reached the intended goals in a positive manner. The global throughput (assuming a distributed volume of transactions) is 5.43 messages per second with one instance, and 20.17 messages per second with five instances.

Table 5.5: Average processing time and combined throughput.

Functionality	**NLP Pipeline**	**Processing Time (seconds)**	**Throughput (messages per second)**	
			1 instance	5 instances
Request Permission	✗	0.03	5.43	20.17 (4.03)
Get PII Types	✗	0.03		
Get PII (DS to DP)	✓	0.58		
Get PII (DP to DP)	✓	0.59		
Update PII	✓	1.14		
PII Analysis (DS)	✓	1.14		
PII Analysis (DP)	✓	1.12		
Registration Notification	✗	0.02		

The proposed solution is currently being evaluated under the PoSeID-on's Use Cases in four distinct pilots: (1) Ministry of Economy and Finance, Italy; (2) Softeam, France; (3) Santander Municipality, Spain; and (4) Malta Information and Technology Agency, Malta. The feedback collected so far indicates not only a complete and successful integration, but also generally positive feedback from the end-users involved in the Use Cases.

The next section provides a summary of the chapter, highlighting the main contributions and validation results.

5.4 Summary

This chapter proposed two privacy risk assessment methods. These mechanisms allow two different types of analysis: a crisp model for binary classification of privacy risk, and a fuzzy model enabling the quantification of privacy risks. The risk assessment is made possible by analysing a combination of inputs: type of data, size, duration, reputation, and data sensitiveness.

The privacy risk assessment mechanism was deployed in the PoSeID-on platform and validated under real-world circumstances. The integration allowed the assessment of the method and its performance with real applicability and benefit of data subjects. Moreover, the output of the mechanism enables the generation of privacy warnings whenever privacy risk thresholds are crossed. The user notification approach has the potential to avoid or minimise privacy exposures by raising users' awareness of their data and risks.

The outcomes of this chapter resulted in the following paper:

- Silva P., Gonçalves C., Godinho C., Antunes N. and Curado M. (2021). **Privacy Risk Assessment and Privacy-preserving Data Monitoring**", IEEE Access, 2021 (submitted, under review).

The following chapter presents a synthesis of this thesis and its main contributions, as well as future work directions.

Chapter 6

Conclusions and Future Work

Contents

Privacy risks are increasingly difficult to predict due to the growing number of personally identifiable information produced, collected and shared among users and companies. The exchanged information is useful for a variety of situations, such as legitimately driving business or accessing services and products. However, despite regulations like GDPR, hundreds of privacy breaches occur every year, either due to inappropriate usage of user data or due to security issues exploited by third parties. Therefore, preventing abusive situations and minimising privacy-compromising scenarios is of the utmost importance.

This thesis proposed several mechanisms that improve privacy assurances in Cloud environments from the perspectives of secure data access, privacy-preserving automated data analysis and privacy risk estimation. This chapter reviews the main contributions of this thesis and addresses future research directions.

6.1 Synthesis of the Thesis

Safekeeping users' private information in the Cloud, raising privacy awareness and reducing privacy risks are the main goals behind the work developed in this thesis. For that, a combination of security and privacy mechanisms are proposed to accomplish those goals and minimise privacy risks in the Cloud. The process included the definition and development of an AAAaaS module, a hybrid data analysis approach that recurs to NLP tools and ML models to monitor and classify PII, and crisp and fuzzy models that assess privacy risks.

Chapter 2 presented the research background and introduced the motivation for this thesis. Specifically, it provided background on PETs, introduced various privacy concepts, technologies, threats and regulations. Finally, it discussed specific data protection and privacy challenges in the Cloud.

Chapter 3 proposed AAAaaS, a cloud-based security module that was deployed and successfully validated in the EUBra-BIGSEA Cloud platform. The chapter also describes how the module integrated with other components to serve different applications and services. Moreover, details of the web interface and its REST API specifically designed for authentication, authorisation and accounting are also described.

Chapter 4 presented different approaches for data analysis using NLP, NER and custom-made ML models. This culminates in the presentation of a hybrid classification mechanism that allows the automation of data analysis tasks, while maintaining the characteristics of a privacy-preserving system. The proposed approach performed better than similar mechanisms in the correct identification of PII. Furthermore, the final section of the chapter presents the lessons learned and how the proposed approach can effectively act as a Privacy Enhancing Technology.

Chapter 5 proposed two privacy risk assessment methods that analyse privacy risks based on common transaction elements such as data types, amount, dura-

tion of processing, reputation, and data sensitiveness. These mechanisms offered two types of models: a crisp model for binary classification of privacy risk and a fuzzy model that enables the quantification of privacy risks.

Moreover, Chapter 5 also described how the privacy risk assessment mechanisms and the hybrid data analysis mechanisms proposed in Chapter 4 resulted in a Cloud module denominated *Personal Data Analyser (PDA)*. It also analysed the module's validation in PoSeID-on's Cloud platform, which allowed the generation of privacy warnings whenever privacy risks were detected and thus raised users' awareness regarding their data and associated risks.

6.2 Contributions

The main objectives of this thesis were introduced in Chapter 1 and are shortly revisited in this section.

One objective was to devise a cloud-based mechanism for providing AAA functionalities as a service.

Another objective was to propose AI mechanisms for automated personal data analysis and classification. These mechanisms aimed to provide a compliant data analysis solution that can be used for the sole benefit of users.

The third objective was to propose privacy risk assessment mechanisms that operate in Cloud environments and process multiple inputs.

The final and more general objective was to validate the proposed mechanism in real Cloud platforms with real users.

These objectives have led to the work presented in this thesis and have resulted in the following contributions:

Contribution 1, Design and development of an AAA mechanism

This contribution, presented in Chapter 3, focused on security aspects of data access. It refers to the design and development of an elastic and efficient AAA security module specifically designed for the Cloud. It provides support to both infrastructure and applications. The possibility of interfacing with external identity providers also enhanced the potential of this service, as seen during integration and validation (Section 3.3).

Contribution 2, Named Entity Recognition models for PII classification

Chapter 4 presented several contributions of this Thesis. Among them is a method based on ML algorithms such as MLP and RF for privacy violation detection through the identification of personal information. This contribution, presented in Section 4.4, involved training word vectors, using as pre-trained word vectors, using publicly available data sources, labelling entities, training, and evaluating models. The proposed models achieved high $F_1 scores$ in the classification of named entities associated to PII. Prior to this work, a comparative study of existent NLP

tools was also performed. The results, described in Section 4.3, offered evidence for the inclusion of some of the tools in a hybrid classification mechanism (Section 4.5).

Contribution 3, Hybrid data classification mechanism

Another contribution included in Chapter 4 was a hybrid NER pipeline for detection of personally identifiable information (described in Section 4.5). This mechanism included the results from Contribution 2 as well as NLP tools and specifically tailored regular expressions. The outcome was an ensemble mechanism capable of collecting the output from all the classifiers and inferring the entity that best matched the input data. The aim of the proposed mechanism was not to provide a perfect data analysis solution - a challenge in NLP - but rather a practical way of handling sensitive data in an autonomous, effective and privacy-preserving manner.

Contribution 4, Multi-input privacy risk assessment mechanism

Chapter 5 showcased multi-input mechanisms built to assess privacy risks in data transactions in crisp and fuzzy perspectives. The models' inputs, described in Section 5.2, included data sensitiveness, correlation, retention time, data validation, and reputation of the parties involved in data transactions. The crisp privacy risk model, described in Section 5.2.4, provides a binary classification of privacy risks. The fuzzy logic approach, described in Section 5.2.5, provided a model capable of offering a normalised quantification of privacy risk levels associated with data transactions.

Contribution 5, Automatic mechanism for privacy-enhancing data analysis

Chapter 5 presented a privacy-enhancing data analysis mechanism. The design and development of the mechanism, capable of automatic data analysis, was made possible with the mechanisms proposed in Contributions 2, 3 and 4. The proposed mechanism leverages a combination of data analysis and privacy risk assessment techniques to compute privacy risks for the sole benefit of data subjects. Then, it generates different levels of privacy warnings capable of raising user awareness and minimising privacy threats.

This contribution resulted in a software denominated *Personal Data Analyser (PDA)*. The PDA was built to be deployed in Cloud platforms and monitor data transactions every time a user grants it permission to perform data analysis. Its main objectives are performing privacy-preserving data analysis and issuing privacy risk notifications directly to users of the platforms where it is deployed.

Contribution 6, Impact assessment of the proposed mechanisms in real-world Cloud platforms

The methods and mechanisms proposed in this thesis were deployed,

integrated and validated in Cloud platforms. Contribution 1 was deployed, integrated and validated in EUBra BIG-SEA's Cloud platform. Similarly, Contribution 5 was also deployed, integrated and validated in PoSeID-on's Cloud platform. Both AAAaaS (Contribution 1) and PDA (Contribution 5) were validated by real users of applications and services running in the platforms. Overall, the pilot testing provided positive feedback across the board, demonstrating the real impact the proposed mechanisms have on Cloud platforms and its users.

6.3 Future Work

Despite the positive and promising outcomes of the mechanisms proposed in this thesis, there are a few aspects that can be addressed in future work to enhance the proposed mechanisms. One of the challenges faced in the experimental work was the difficulty in obtaining appropriate data for ML model training. It was quite hard due to the lack of publicly available PII datasets and the shortage of annotated data (in the few available datasets). Future work featuring Federated Learning (FL) may positively contribute to minimise data sourcing limitations. FL enables collaborative model training in a privacy-preserving manner by training models on different servers or devices, in a distributed fashion. Thus, sensitive data never leaves the owners' premises or device - rendering valuable data for the models and providing higher privacy assurances.

Many data processing requests are accompanied by the reasons or purposes of the request. From a privacy point of view, such information is fundamental and should be useful for every user. However, it is rarely fully read or adequately understood by the majority of people. The mechanisms proposed in this thesis apply NLP mechanism such as NER to analyse and classify data. There are other NLP tasks such as NLU that can be useful to analyse and interpret the reasons behind data requests automatically. Nevertheless, this is a challenging task, and its applicability can be limited to specific fields or services.

Another advantage of our system is for permission checking purposes. In this case, permission-based systems would be able to map and verify if the actual data matches the textual description of the respective permissions granted. However, in this particular situation, it would be necessary to devise and implement an NLU module for the extraction of the meaning of such permissions. In systems where there no such textual descriptions of the permissions, it is possible to directly map the permission type to the PII type, thus allowing permission verification on a higher level.

Similarly, privacy policies of data processors can also be a target of automated analysis and respective privacy risk analysis. Mechanisms able to automatically process and interpret privacy policies in machine-readable formats are not new. The literature shows contributions in this area, and there are commercial solutions capable of performing such tasks. However, taking advantage of such capabilities and performing privacy risk analysis considering privacy policies' information could greatly benefit the privacy-enhancing systems.

References

Al Omran, F. N. A. and Treude, C. (2017). Choosing an nlp library for analyzing software documentation: a systematic literature review and a series of experiments. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 187–197. IEEE.

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Alic, A. S., Almeida, J., Aloisio, G., Andrade, N., Antunes, N., Ardagna, D., Badia, R. M., Basso, T., Blanquer, I., Braz, T., Brito, A., Elia, D., Fiore, S., Guedes, D., Lattuada, M., Lezzi, D., Maciel, M., Meira, W., Mestre, D., Moraes, R., Morais, F., Pires, C. E., Kozievitch, N. P., dos Santos, W., Silva, P., and Vieira, M. (2019). Bigsea: A big data analytics platform for public transportation information. *Future Generation Computer Systems*, 96:243 – 269.

Alqubaisi, F., Wazan, A. S., Ahmad, L., and Chadwick, D. W. (2020). Should we rush to implement password-less single factor fido2 based authentication? In *2020 12th Annual Undergraduate Research Conference on Applied Computing (URC)*, pages 1–6.

Amazon (2012). AWS Identity and Access Management. `https://docs.amazonaws.cn/en_us/IAM/latest/APIReference/iam-api.pdf`.

Apache Mesos (2020). Mesos- A distributed systems kernel. `http://mesos.apache.org`.

Armando, A., Carbone, R., Compagna, L., Cuellar, J., and Tobarra, L. (2008). Formal analysis of saml 2.0 web browser single sign-on: breaking the saml-based single sign-on for google apps. In *Proceedings of the 6th ACM workshop on Formal methods in security engineering*, pages 1–10. ACM.

Athena Tech (2019). AI, Machine Learning (ML) and Natural Language Processing (NLP). `https://athenatech.tech/blog/f/ai-machine-learning-ml-and-natural-language-processing-nlp`.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, 3:19–48.

Barth, A., Datta, A., Mitchell, J. C., and Nissenbaum, H. (2006). Privacy and contextual integrity: Framework and applications. In *Proceedings of the 2006*

IEEE Symposium on Security and Privacy, SP '06, pages 184–198, Washington, DC, USA. IEEE Computer Society.

Barua, A., Mudunuri, L. S., and Kosheleva, O. (2013). Why trapezoidal and triangular membership functions work so well: Towards a theoretical explanation.

Bellovin, S., Schiller, J., and Kaufman, C. (2003). Security mechanisms for the internet. *Request For Comments 3631 (RFC3631)*.

Bing, C. (2020). Suspected russian hackers spied on u.s. treasury emails - sources. `https://www.reuters.com/article/BigStory12/idUSKBN28N0PG`.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Bonneau, J., Herley, C., Van Oorschot, P. C., and Stajano, F. (2012). The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy*, pages 553–567. IEEE.

Burns, A. and Johnson, E. (2018). The evolving cyberthreat to privacy. *IT Professional*, 20(3):64–72.

Butler, D. (2007). Data sharing threatens privacy. *Nature*, 449(7163):644–646.

Canale, L., Lisena, P., and Troncy, R. (2018). A novel ensemble method for named entity recognition and disambiguation based on neural network. In Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M. C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E., editors, *The Semantic Web – ISWC 2018*, pages 91–107, Cham. Springer International Publishing.

Cantor, S., Moreh, J., Philpott, R., and Maler, E. (2004). Metadata for the oasis security assertion markup language (saml) v2. 0. *OASIS Standard (March 2005)*.

Carrel, D. and Grant, L. (1997). Cisco systems, the tacacs+ protocol. Technical report, IETF.

Chana, I. and Singh, S. (2014). Quality of service and service level agreements for cloud environments: Issues and challenges. In *Cloud Computing*, pages 51–72. Springer.

Chen, J., Huang, H., Tian, S., and Qu, Y. (2009). Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, 36(3):5432–5435.

Chilberto, J., Zaal, S., Aroraa, G., and Price, E. (2020). Identity security with azure active directory. In *Cloud Debugging and Profiling in Microsoft Azure*, pages 215–234. Springer.

Choudhury, A. J., Kumar, P., Sain, M., Lim, H., and Jae-Lee, H. (2011). A strong user authentication framework for cloud computing. In *2011 IEEE Asia-Pacific Services Computing Conference*, pages 110–115. IEEE.

Cimpanu, C. (2020). Microsoft discloses security breach of customer support database. `https://www.zdnet.com/article/microsoft-discloses-security-breach-of-customer-support-database/`.

Cloudflare, Inc. (2020). Cloudflare ssl/tls. `https://www.cloudflare.com/en-gb/ssl/`.

Coron, J.-S. (2006). What is cryptography? *IEEE security & privacy*, 4(1):70–73.

Cragin, M., Heidorn, B., Palmer, C., and Smith, L. (2007). An educational program on data curation. `http://hdl.handle.net/2142/3493`.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., and Wilks, Y. (2000). Experience using gate for nlp r&d. In *Proceedings of the COLING-2000 Workshop on Using Toolsets and Architectures To Build NLP Systems*, pages 1–8.

Daintith, J. (2019). A dictionary of computing. 2004. `http://www.encyclopedia.com/doc/1O11-anonymization.html`.

De Joyee, S. and Le Métayer, D. (2016). Priam: A privacy risk analysis methodology. In Livraga, G., Torra, V., Aldini, A., Martinelli, F., and Suri, N., editors, *Data Privacy Management and Security Assurance*, pages 221–229, Cham. Springer International Publishing.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhasarathan, C., Thirumal, V., and Ponnurangam, D. (2015). Data privacy breach prevention framework for the cloud service. *Security and Communication Networks*, 8(6):982–1005.

Domingo-Ferrer, J., Farràs, O., Ribes-González, J., and Sánchez, D. (2019). Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges.

Drake, J. (2019). Four types of invasion of privacy. `https://legalbeagle.com/8068982-four-types-invasion-privacy.html`.

Dunn, J. (2018). Uber data breach aided by lack of multi-factor authentication. `https://nakedsecurity.sophos.com/2018/02/08/uber-data-breach-aided-by-multi-factor-authentication-weakness/`.

ENISA (2009). Benefits, risks and recommendations for information security. *European Network and Information Security*.

Explosion AI (2020). English - spacy models documentation. https://spacy.io/models/en.

ExplosionAI (2020). spaCy - Industrial-Strength Natural Language Processing. https://spacy.io.

Fajardo, V., Arkko, J., Loughney, J., and Zorn, G. (2012). Diameter base protocol. *IETF (The Internet Engineering Task Force) Request for Comments*, 6733.

Federal Deposit Insurance Corporation (2019). FDIC: Privacy Act Issues under Gramm-Leach-Bliley. https://www.fdic.gov/consumers/consumer/alerts/glba.html.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53.

Gachanga, E., Kimwele, M., and Nderu, L. (2018). Sensitivity based anonymization with multi-dimensional mixed generalization. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 168–172, Piscataway, NJ, USA. IEEE.

Gibbs, S. (2015). The guardian - what is 'safe harbour' and why did the eucj just declare it invalid? https://www.theguardian.com/technology/2015/oct/06/safe-harbour-european-court-declare-invalid-data-protection.

Google (2019). Syntaxnet: Neural models of syntax. https://github.com/tensorflow/models/tree/master/research/syntaxnet.

Google Code Archive (2020). word2vec. https://code.google.com/archive/p/word2vec/.

Greenleaf, G. and Chen, H.-l. (2012). Data Privacy Enforcement in Taiwan, Macau, and China. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2118332.

Grosso, A. J., Leffard, G. W., and O'dwyer, J. G. (2014). System and method for analyzing privacy breach risk data. US Patent App. 13/683,422.

Hankerson, D., Menezes, A. J., and Vanstone, S. (2006). *Guide to elliptic curve cryptography*. Springer Science & Business Media.

Hardt, D. et al. (2012). The oauth 2.0 authorization framework. Technical report, RFC 6749, October.

Heidari, P., Lemieux, Y., and Shami, A. (2016). Qos assurance with light virtualization-a survey. In *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 558–563. IEEE.

Hickman, K. and Elgamal, T. (1995). The ssl protocol. Technical report, Netscape Communications Corp.

Housley, R., Polk, W., Ford, W., and Solo, D. (2002). Rfc3280: Internet x. 509 public key infrastructure certificate and certificate revocation list (crl) profile.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification.

Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59.

ISO (2009). Plastics – Determination of fracture toughness – Linear elastic fracture mechanics (LEFM) approach. Standard, International Organization for Standardization, Geneva, CH.

Jasti, A., Shah, P., Nagaraj, R., and Pendse, R. (2010). Security in multitenancy cloud. In *44th Annual 2010 IEEE International Carnahan Conference on Security Technology*, pages 35–41. IEEE.

Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27.

Jindal, N. and Liu, B. (2007). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190.

Kaggle Inc. (2020). An online community of data scientists and machine learners. `https://www.kaggle.com`.

Khalid, U., Ghafoor, A., Irum, M., and Shibli, M. A. (2013). Cloud based secure and privacy enhanced authentication & authorization protocol. *Procedia Computer Science*, 22:680–688.

Khalil, I. M., Khreishah, A., Bouktif, S., and Ahmad, A. (2013). Security concerns in cloud computing. In *2013 10th International Conference on Information Technology: New Generations*, pages 411–416.

Kirk, J. (2014). Jpmorgan chase says breach affected 83m customers. `https://www.computerworld.com/article/2691246/jpmorgan-chase-says-breach-affected-83m-customers.html`.

Krishnamurthy, B. and Wills, C. E. (2009). On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, pages 7–12, New York, NY, USA. ACM.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lambert, J. (2020). Important steps for customers to protect themselves from recent nation-state cyberattacks. `https://blogs.microsoft.com/on-the-issues/2020/12/13/customers-protect-nation-state-cyberattacks/`.

Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., and Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36(1):16 – 24.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

López, D. (2006). eduGAIN: Federation interoperation by design. In *TERENA Networking Conference*.

Loria, S. (2019). Textblob documentation. `https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf`.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.

Mercuri, R. T. (2004). The hipaa-potamus in health care data security. *Commun. ACM*, 47(7):25–28.

Mesosphere, Inc. (2020). Marathon. `https://mesosphere.github.io/marathon/`.

Metrolink (2019). Southern california regional rail authority. `https://www.metrolinktrains.com/globalassets/about/contracts/may-26-2019/contract-no.-sp452-16-conformed-contract-fully-executed.pdf`.

Microsoft (2014). Protecting data and privacy in the cloud. Technical report, Microsoft. `http://download.microsoft.com/download/2/0/a/20a1529e-65cb-4266-8651-1b57b0e42daa/protecting-data-and-privacy-in-the-cloud.pdf`.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Milinovic, M. et al. (2008). Eduroam: Service definition and implementation plan. *GEANT2, Deliverable DS5*, 1.

Milne, G. R., Pettinico, G., Hajjat, F. M., and Markos, E. (2017). Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *Journal of Consumer Affairs*, 51(1):133–161.

Moon, M. (2018). President signs overseas data access bill into law. `https://www.engadget.com/2018/03/24/cloud-act-law/`.

MostlyAI (2020). Creating AI-generated synthetic data. `https://mostly.ai`.

Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105:24.

N.C. Board of Elections (2020). North carolina state board of elections - election results data. `https://www.ncsbe.gov/Public-Records-Data-Info/Election-Results-Data`.

Neves, M. and Ševa, J. (2019). An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*. bbz130.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, 50(3):559–569.

Nuñez, D. and Agudo, I. (2014). Blindidm: A privacy-preserving approach for identity management as a service. *International Journal of Information Security*, 13(2):199–215.

OpenIdentityPlatform (2020). OpenAM. `https://github.com/OpenIdentityPlatform/OpenAM`.

OpenStack Foundation (2020). Keystone, the openstack identity service. `https://docs.openstack.org/keystone/latest/`.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083.

Peisert, S., Talbot, E., and Kroeger, T. (2013). Principles of authentication. In *Proceedings of the 2013 New Security Paradigms Workshop*, pages 47–56.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perkins, C. and Calhoun, P. (2005). Authentication, authorization, and accounting (aaa) registration keys for mobile ipv4. Technical report, RFC 3957, March.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Petitcolas, F. A., Anderson, R. J., and Kuhn, M. G. (1999). Information hiding-a survey. *Proceedings of the IEEE*, 87(7):1062–1078.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Recordon, D. and Reed, D. (2006). Openid 2.0: a platform for user-centric identity management. In *Proceedings of the second ACM workshop on Digital identity management*, pages 11–16.

Rigney, C., Willens, S., Rubens, A., and Simpson, W. (2000). Remote authentication dial in user service (radius).

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534.

Rivest, R. L., Shamir, A., and Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126.

Robison, W. L. (2018). *Digitizing Privacy*, pages 189–204. Springer International Publishing, Cham.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., and Suter, B. W. (1990). The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, 1(4):296–298.

Sadollah, A. (2018). *Fuzzy Logic Based in Optimization Methods and Control Systems and Its Applications*. BoD–Books on Demand.

Saripalli, P. and Walters, B. (2010). Quirc: A quantitative impact and risk assessment framework for cloud security. In *2010 IEEE 3rd international conference on cloud computing*, pages 280–288. Ieee.

Schneier, B. (1999). The uses and abuses of biometrics. *Communications of the ACM*, 42(8):136–136.

Schonrock, J. et al. (2018). Invasion of Privacy. Technical report, Find Law - Thomson Reuters.

Schulz, M. and Hennis-Plasschaert, J. A. (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. *Official Journal of the European Union*, 59:88.

Senate of the United States (2018). Clarifying Lawful Overseas Use of Data Act or the "CLOUD Act". `https://www.hatch.senate.gov/public/_cache/files/6ba62ebd-52ca-4cf8-9bd0-818a953448f7/ALB18102(1).pdf`.

Sergey, M. (2018). Data protection in Russian Federation: overview - Practical Law. `https://uk.practicallaw.thomsonreuters.com/2-502-2227?__lrTS=20180419130106547&transitionType=Default&contextData=(sc.Default)`.

Silva, P., Basso, T., Antunes, N., Moraes, R., Vieira, M., Simoes, P., and Monteiro, E. (2018). A europe-brazil context for secure data analytics in the cloud. *IEEE Security Privacy*, 16(6):52–60.

Singh, A. and Chatterjee, K. (2017). Cloud security issues and challenges: A survey. *Journal of Network and Computer Applications*, 79:88–115.

Solove, D. (2009). *Understanding Privacy*, volume 173. Harvard University Press, MA, USA.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Tang, J., Cui, Y., Li, Q., Ren, K., Liu, J., and Buyya, R. (2016). Ensuring security and privacy preservation for cloud data services. *ACM Comput. Surv.*, 49(1).

Tari, Z., Yi, X., Premarathne, U. S., Bertok, P., and Khalil, I. (2015). Security and privacy in cloud computing: Vision, trends, and challenges. *IEEE Cloud Computing*, 2(2):30–38.

Tavares, H. L., Rezende, G. G., dos Santos, V. M., Manhães, R. S., and de Carvalho, R. A. (2010). A tool stack for implementing behaviour-driven development in python language. *CoRR*, abs/1007.1722.

TensorFlow (2019). An end-to-end open source machine learning platform. `https://www.tensorflow.org`.

Texas DoIR (2019). Texas department of information resources - our mission is to provide technology leadership, technology solutions. `https://dir.texas.gov/View-Search/Contracts-Detail.aspx?contractnumber=DIR-TSO-4101`.

The Linux Foundation (2020). Kubernetes - Production-Grade Container Orchestration. https://kubernetes.io.

Todd, S., Baldwin, R., Dietrich, D., and Pauley Jr, W. A. (2016). Privacy scoring for cloud services. US Patent 9,356,961.

Tomashchuk, O., Van Landuyt, D., Pletea, D., Wuyts, K., and Joosen, W. (2019). A data utility-driven benchmark for de-identification methods. In Gritzalis, S., Weippl, E. R., Katsikas, S. K., Anderst-Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Trust, Privacy and Security in Digital Business*, pages 63–77, Cham. Springer International Publishing.

Trautman, L. J. and Ormerod, P. C. (2016). Corporate directors' and officers' cybersecurity standard of care: The yahoo data breach. *Am. UL Rev.*, 66:1231.

University of Groningen (2019). Groningen meaning bank. https://gmb.let.rug.nl.

UpGuard (2019). Losing Face: Two More Cases of Third-Party Facebook App Data Exposure. https://www.upguard.com/breaches/facebook-user-data-leak.

US Code (1999). Title V of the Gramm-Leach-Bliley Act's (GLBA). https://www.gpo.gov/fdsys/pkg/USCODE-2011-title15/pdf/USCODE-2011-title15-chap94-subchapI.pdf.

U.S. DoC (2000). U.S. Department of Commerce - Safe Harbor Privacy Principles. https://2016.export.gov/safeharbor/eu/eg_main_018475.asp.

U.S. DoD (2019). Official website for u.s. department of defense. https://www.defense.gov/Newsroom/Contracts.

Vavilis, S., Petković, M., and Zannone, N. (2014). A reference model for reputation systems. *Decision Support Systems*, 61:147–154.

Vlachos, A. (2007). Evaluating and combining and biomedical named entity recognition systems. In *Biological, translational, and clinical language processing*, pages 199–200.

Wang, H., Liu, F., and Liu, H. (2012). A method of the cloud computing security management risk assessment. In *Advances in Computer Science and Engineering*, pages 609–618. Springer.

Wiles, R., Crow, G., Heath, S., and Charles, V. (2008). The management of confidentiality and anonymity in social research. *International Journal of Social Research Methodology*, 11(5):417–428.

Woodward, J. D. (1997). Biometrics: privacy's foe or privacy's friend? *Proceedings of the IEEE*, 85(9):1480–1492.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yoo, Y. (2010). Computing in everyday life: A call for research on experiential computing. *MIS quarterly*, pages 213–231.

Yun, H., Lee, G., and Kim, D. J. (2019). A chronological review of empirical research on personal information privacy concerns: An analysis of contexts and research constructs. *Information & Management*, 56(4):570 – 601.

Zaman, A., Obimbo, C., and Dara, R. A. (2017). An improved differential privacy algorithm to protect re-identification of data. In *IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 133–138, Toronto, Canada. IEEE.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

Zhang, L. and Suganthan, P. N. (2016). A survey of randomized algorithms for training neural networks. *Information Sciences*, 364:146–155.