

Article

A Hybrid Approach to the Automatic Detection of Personal Data in Latvian-Language Texts

Henrihs Gorskis ^{1,*}, Jūlija Strebko ^{1,*}, Jurijs Korņijenko ², Vitālijs Zabiņako ² and Andrejs Romanovs ^{1,*}¹ Faculty of Computer Science, Information Technology and Energy, Riga Technical University, Kipsalas iela 6A, LV-1048 Rīga, Latvia² SIA "ABC Software", Tallinas iela 51A, LV-1012 Rīga, Latvia

* Correspondence: henrihs.gorskis@rtu.lv (H.G.); julija.strebko@rtu.lv (J.S.); andrejs.romanovs@rtu.lv (A.R.)

Abstract

In the modern world, hybrid and fully remote work formats are becoming increasingly widespread. The volume of digital communication continues to grow, and the need to exchange documents and information through email, corporate messengers (such as Microsoft Teams), and collaborative workspaces is rising. Personal data is often involved in these exchanges, which increases the risk of unintentional disclosure. To comply with GDPR requirements and ensure information security, it is necessary to implement methods for the automatic detection of personal data. This task is particularly relevant for low-resource languages, such as Latvian, for which existing tools often operate with limited accuracy and efficiency. This work proposes a hybrid approach to the automatic detection of personal data in Latvian texts from Microsoft Teams messages, emails, and documents, combining a transformer-based NER model with rule-based detection of structured identifiers. The approach builds on a multilingual NER model, supplements it with Latvian-specific rules and structured-identifier detectors, and demonstrates the potential of adapted solutions to improve the accuracy and robustness of personal-data detection in real-world workflows.

Keywords: GDPR; personal data; confidentiality classification; PII detection; recognition methods; NLP; NER; information security; risk assessment

1. Introduction

With the development of smart technologies and the extensive use of large-scale cloud services, the volume of personal data circulating within information systems continues to grow. These data constitute a key resource for businesses, government institutions, and research organizations. However, they also create significant privacy risks. Although several security standards and regulations exist, such as the GDPR and the ISO/IEC 27000 series, organizations still lack a universally adopted system for personal-data classification. This gap reduces the effectiveness of security measures and complicates the implementation of automated tools designed to detect and anonymize personal data. Another challenge hindering the use of such tools is the need to process texts in local languages and recognize country-specific identifiers, which vary considerably across countries. Existing solutions demonstrate high effectiveness primarily for the English language, highlighting the need to adapt them for low-resource languages, such as Latvian.

The aim of this study is to develop a hybrid approach for the detection and classification of personal data in Latvian-language texts. To achieve this aim, the following tasks are defined:



Academic Editor: Francesco Buccafurri

Received: 30 May 2026

Revised: 30 June 2026

Accepted: 7 July 2026

Published: 9 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Analyze the relevant scientific literature;
2. Examine existing approaches to personal-data recognition and classification, and identify their limitations;
3. Define detectors for Latvia-specific identifiers;
4. Implement and evaluate a hybrid approach to personal-data detection and classification in Latvian-language texts.

The work described in this paper is done in partnership with “ABC software” SIA (LLC). The research leading to these results has received funding from the project “Competence Centre of Information and Communication Technologies for Digitalization” of the Recovery and Resilience Facility, contract No. 2.2.1.3.i.0/1/24/A/CFLA/006 signed between IT Competence Centre and Central Finance and Contracts Agency, Research No. 1.1 “Analysis, Evaluation, and Classification of User-Generated Digital Content in IT Systems Using AI/ML Approaches”.

2. Background and Related Work

Automatic detection of personal data in texts is an increasingly important task as the volume of digital communication grows and data protection requirements become more stringent. Various methods can be applied to this task, including regular expressions, dictionary-based methods, transformer-based models, and context-aware models. However, the effectiveness of these methods depends strongly on the language in which the texts are written and may decrease when they are applied to low-resource languages. This section reviews relevant regulatory frameworks, methodological approaches, and existing tools and platforms for the automatic detection and classification of personal data.

2.1. International Standards and Frameworks

International standards and regulatory frameworks establish a methodological basis for the classification, processing, and protection of personal data. They define fundamental principles for information management, security, and privacy protection.

The most important regulatory document for the protection of personal data in the European Union is the General Data Protection Regulation (GDPR), which applies across the EU Member States [1]. It defines personal data broadly and distinguishes special categories of personal data that require enhanced protection. The GDPR also requires a Data Protection Impact Assessment (DPIA) where processing is likely to result in a high risk to the rights and freedoms of natural persons [2]. However, the GDPR does not provide a formal scale for determining the sensitivity level of personal data. Instead, it establishes principles according to which organizations must assess risks and implement appropriate data protection measures. Similar regulations also exist in other jurisdictions, such as the California Consumer Privacy Act (CCPA) in California, USA [3], and the Lei Geral de Proteção de Dados (LGPD) in Brazil [4].

The ISO/IEC 27000 series of standards provides a framework for information security management. These standards do not define specific categories of personal data. Rather, they establish a methodological basis on which organizations can develop their own information classification schemes and protection processes. Three standards within this family are particularly relevant to the processing and protection of personal data:

- ISO/IEC 27001 specifies requirements for the establishment, implementation, maintenance, and continual improvement of an information security management system (ISMS) [5].
- ISO/IEC 27002 provides guidance on information security controls, including the classification of information (public, internal, confidential, restricted) according to its sensitivity and importance to the organization [6].

- ISO/IEC 27701 specifies requirements and provides guidance for establishing, implementing, maintaining, and continually improving a privacy information management system (PIMS). It addresses the processing of personally identifiable information (PII), including the roles of PII controllers and PII processors, and can be aligned with an existing ISO/IEC 27001-based ISMS [7].

In this paper, the term personal data is used in accordance with the GDPR. The term personally identifiable information (PII) is retained where it is used in the terminology of ISO/IEC standards or the documentation of specific technological tools.

2.2. Approaches and Methods for Personal-Data Classification and Detection

Under the GDPR, personal data are any information relating to an identified or identifiable natural person who can be identified, directly or indirectly, by reference to information such as a name, an identification number, location data, or another identifying characteristic [1].

Effective classification of personal data is based on a combination of legal principles, contextual analysis, and technical methods. Personal data classification commonly considers the following dimensions:

- Distinguishing among general personal data, special categories of personal data, such as health or biometric data, and other potentially high-risk identifiers, such as personal identification numbers or financial identifiers;
- Assessing the risks of disclosure and the possibility of identifying or re-identifying individuals;
- Determining requirements for the storage and processing of personal data.

2.2.1. Classification of Personal Data

The classification of personal data is based on criteria that are largely determined by legal and regulatory frameworks, particularly the GDPR. This regulation establishes the principles of lawfulness, data minimization, and purpose limitation, requiring that personal data be adequate, relevant, and limited to what is necessary for the purposes for which they are processed. These principles provide a basis for distinguishing data categories, determining their sensitivity, and assessing the risks of re-identification.

Research indicates that data classification in different jurisdictions is based on distinctions between general personal data, special categories of personal data, and other data types requiring enhanced protection. For example, Supriyadia et al. (2023) indicate that, in Indonesia, personal data include names, addresses, identification numbers, financial information, and medical information, which require differentiated approaches to their processing [8]. Similarly, an analysis of legislation in the United Arab Emirates shows that the protection of electronic personal data is supported through storage requirements, access control mechanisms, and civil liability provisions [9].

In the context of e-commerce, data classification is considered an essential prerequisite for ensuring user trust. This is especially important because online platforms process large volumes of data, including transactional, behavioral, and identification data, which require classification according to their sensitivity levels [10].

Thus, regulatory documents provide the basis for classification schemes by determining which data require protection and which criteria should be considered in their processing. In general, personal data may be differentiated into general personal data, special categories of personal data, and other data types that may create elevated risks in a particular processing context. Depending on the application scenario, classification schemes may additionally distinguish categories such as identification, financial, medical, transactional, behavioral, and biometric data.

2.2.2. Risk-Based Assessment Methods

Risk-based approaches enable the assessment of the likelihood of unauthorized disclosure of personal data and its possible consequences. Several studies emphasize that personal-data classification should consider not only the type of data involved, but also the potential risks in the event of a data breach. For example, Utami et al. (2025) analyzed actual data breach cases in Indonesia and indicated that medical, financial, and biometric data require particularly strict protection because of their potential for misuse [11]. The study identified medical, financial, biometric, and identity-card data as high-risk data categories. Potential consequences include fraud, identity theft, discrimination, and financial loss.

In international practice, risk-based approaches are also reflected in information security standards. Previous research has examined how the information security controls defined in ISO/IEC 27002 can be extended or adapted to support GDPR compliance, including the assessment of re-identification risks and the implementation of additional protection measures [12]. In such an approach, data can be classified according to risk level, for example, as low, medium, or high risk, enabling potential threats to be associated with appropriate protection measures. Attention should be paid to the risk of re-identification, especially where special categories of personal data or other high-risk identifiers are processed. In such cases, risk assessment may indicate the need for additional technical and organizational measures, supporting a differentiated approach to the protection of different categories of information.

2.2.3. Named Entity Recognition (NER) for Personal-Data Detection

NER models are widely used automated methods for detecting entities that may represent personal data in texts. Multilingual transformer models, such as XLM-R, can identify entities including personal names, addresses, organization names, and geographic locations. Some of these entity categories, particularly personal names and addresses, may directly correspond to personal data, while others may contribute to the identification of an individual depending on the context. This makes NER a useful component of personal-data detection and anonymization systems [13].

In some studies, NER models have also been applied to the detection of sensitive information in social media texts. For example, Erol et al. (2022) investigated the detection of disclosed personal health information in Turkish social media data and reported strong classification performance [14]. In clinical contexts, hybrid NER approaches have also been used to extract clinical parameters from free-text medical records [15].

2.2.4. Rule-Based and Regular-Expression Methods

Since some types of personal data have structured formats, such as telephone numbers, email addresses, Latvian personal identification numbers, IBANs, and payment card numbers, regular expressions are a useful method for their detection. By combining regular expressions with specialized dictionaries, structured patterns relevant to personal-data detection can be efficiently identified in large volumes of text [16]. Krishnaram (2025) emphasizes the usefulness of regular expressions for data validation and information extraction, demonstrating their applicability to various structured data types [17].

2.2.5. Context-Aware Approaches

In personal-data detection and classification, context is essential because the sensitivity and significance of a data element may depend on the semantic context in which it occurs. This is particularly important when classifying ambiguous data elements. Tahir et al. (2025) propose a context-aware model for recognizing personal data based on the analysis of situational factors and context graphs [18]. This approach makes it possible to assess the

risk associated with personal data according to their use: for example, a patient's name linked to medical information may present a higher risk than an author's name appearing in a published article. Context-aware approaches provide greater flexibility in classification and enable methods to be adapted to specific scenarios.

2.2.6. Summary

The reviewed approaches indicate that the classification and detection of personal data constitute a multidimensional process based on legal and regulatory frameworks, risk analysis, and technical methods. Regulatory frameworks, especially the GDPR, establish fundamental criteria for the processing and protection of personal data. Risk-based methods enable the assessment of potential consequences and the adjustment of protection measures according to the sensitivity and context of the data. Technical approaches, ranging from NER models to regular expressions and context-aware analysis, provide practical methods for automatic personal-data detection in various scenarios. Taken together, these findings support the use of a hybrid approach that combines NER-based detection, rule-based identification of structured identifiers, and context-sensitive assessment of detected personal data.

2.3. Overview of Existing Tools

A range of tools is available for the detection and classification of personal data, from enterprise cloud platforms to specialized open-source libraries.

2.3.1. Microsoft Purview

Microsoft Purview is an enterprise platform that supports data classification, information protection, and compliance management [19]. It includes Sensitive Information Types (SITs), which are classifiers used to detect predefined categories of sensitive information, such as financial identifiers, government-issued identifiers, and personal identifiers. SITs may rely on pattern matching, supporting evidence, confidence levels, named-entity detection, or Exact Data Match (EDM), depending on the type of classifier used. Match confidence levels, such as low, medium, and high, make it possible to configure the reliability threshold required for a detected match. Microsoft Purview also supports trainable classifiers for identifying particular categories of content, as well as EDM-based SITs for detecting predefined exact values. The platform integrates with Data Loss Prevention (DLP) policies, automatic labeling mechanisms, and Microsoft 365 services, making it suitable for enterprise information-protection workflows.

2.3.2. Microsoft Presidio

Microsoft Presidio is an open-source framework for detecting and anonymizing personally identifiable information (PII) [20]. It uses a hybrid approach in which recognizers can apply regular expressions, context rules, rule-based logic, and NER models to detect PII entities in unstructured text. Presidio includes an Analyzer component for detecting PII entities and an Anonymizer component for applying de-identification operations to detected entities. As an open-source and extensible framework, Presidio can be supplemented with custom recognizers and adapted to additional entity types or languages, including low-resource languages.

2.3.3. Google Cloud Sensitive Data Protection

Google Cloud Sensitive Data Protection uses information types, or infoTypes, to define the categories of sensitive information to be detected, such as names, email addresses, telephone numbers, identification numbers, and payment card numbers [21]. Each infoType is associated with an infoType detector, which determines how instances of that information

type are identified. The platform supports built-in and custom detectors, including regular-expression detectors, dictionary-based detectors, and context-based inspection rules. It can inspect large volumes of data and integrates with other Google Cloud services, making it suitable for organizations that process data within cloud-based infrastructures.

2.3.4. Amazon Macie

Amazon Macie is a data security service that automatically discovers sensitive data in objects stored in Amazon S3 [22]. It uses machine learning and pattern matching to detect categories of sensitive data, including credentials, financial information, personal health information, and personally identifiable information. Macie provides managed data identifiers and supports custom data identifiers for detecting organization-specific patterns. Its automated sensitive-data discovery functionality continually evaluates S3 bucket inventories and analyses selected objects to identify sensitive data and generate corresponding findings. In addition, Macie provides information that can assist organizations in assessing security and privacy risks associated with their S3 data.

2.3.5. IBM Guardium

IBM Guardium Data Protection is a data protection platform focused on discovering and classifying sensitive data, monitoring data access, and supporting compliance activities across heterogeneous data environments [23]. It provides capabilities for monitoring sensitive-data activity, enforcing access policies, and protecting data through mechanisms such as dynamic masking and redaction. IBM Guardium also supports audit workflows and compliance reporting, making it suitable for organizations that process sensitive data across databases, data warehouses, cloud environments, and other data stores.

2.3.6. Collibra Data Intelligence Platform

The Collibra Data Intelligence Platform is a data and AI governance platform that includes capabilities for data privacy management [24]. Its privacy-related functionality supports the discovery of sensitive data, the centralization of privacy controls, risk assessment, and compliance reporting across organizational data environments. Collibra can help organizations understand how sensitive data move across teams, systems, and third parties through data lineage and data-flow visualization capabilities. The platform also supports the automation of privacy and risk-management workflows, helping organizations manage regulatory requirements and apply data-protection policies consistently.

2.3.7. Summary of Existing Tools

The reviewed tools for personal-data detection and classification differ in their scalability, integration capabilities, and adaptability to specific organizational and sectoral requirements. Enterprise platforms, such as Microsoft Purview, Google Cloud Sensitive Data Protection, Amazon Macie, IBM Guardium, and the Collibra Data Intelligence Platform, provide broader functionality for data governance, protection, risk management, policy enforcement, and integration with cloud or enterprise data environments. By contrast, open-source frameworks, such as Microsoft Presidio, provide extensible components for the detection and anonymization of personal data in text.

Although their specific capabilities differ, the reviewed tools demonstrate that personal-data detection and classification are commonly supported through combinations of rule-based methods, predefined detectors, machine learning techniques, contextual information, and organizational protection workflows. This diversity enables organizations to select or adapt solutions according to their infrastructure, regulatory requirements, and data-processing contexts.

2.4. NLP Challenges for Low-Resource Languages

One of the main challenges in the automatic detection and classification of personal data is the processing of low-resource languages. Many modern methods, including NER models, machine learning algorithms, and context-aware models, are developed and evaluated primarily for English or other well-resourced languages. Consequently, their performance may decrease when they are applied to languages for which fewer annotated datasets and specialized linguistic resources are available.

Studies indicate that multilingual transformer models may also exhibit reduced performance when applied to low-resource languages [13]. Errors may occur particularly in the detection of proper nouns, morphologically complex word forms, and infrequent vocabulary. A related problem is described by Erol et al. (2022), who investigated the detection of personal health information in Turkish social media data and noted the challenges associated with identifying such information in informal texts [14].

An additional challenge arises from contextual variation and semantic ambiguity. Context-aware models require not only linguistic information, but also information about the situation in which a data element occurs [18]. Consequently, models may fail to distinguish, for example, a person's name from an organization name or to interpret medically relevant information correctly in a particular context. Such errors can reduce the accuracy of personal-data detection and limit the practical applicability of context-aware methods.

In some jurisdictions, the development of digital data-processing practices and personal-data protection requirements may progress faster than the development of language-specific automation tools. Studies concerning personal-data protection in Tanzania and Indonesia emphasize the importance of effective data-protection mechanisms in digitally developing environments [25–27]. However, the extent to which these challenges can be addressed through automated language-processing tools depends on the availability of suitable linguistic resources, annotated datasets, and locally adapted detection methods. Similar conclusions are also drawn by Arifin et al. (2025) and Sulthanah et al. (2025), indicating that the lack of language resources hinders the implementation of effective data protection mechanisms and limits the practical use of automation tools [26,27].

Regarding hybrid approaches that combine regular expressions and NLP methods for low-resource languages, studies indicate that such approaches perform relatively well in detecting structured data but are less effective when processing free text [28]. Further challenges arise from the limited availability of training data for low-resource languages, which may cause large language models to reproduce or reinforce inaccurate language patterns [29]. Such errors may increase confidentiality risks, as incorrect classification or failure to detect sensitive data can increase the likelihood of unauthorized disclosure.

In summary, improving the detection and classification of personal data in low-resource languages requires methods that are adapted to the linguistic characteristics of the target language and to relevant national identifier formats. A hybrid approach combining deterministic detectors with machine learning methods provides a suitable basis for addressing these requirements in Latvian-language texts.

3. Research Methodology

The research methodology is based on a hybrid approach that combines a transformer-based NER model with regular expressions, dictionary-based methods, and keyword-based context rules to detect different categories of personal data in Latvian-language texts from Microsoft Teams messages, emails, and documents. This hybrid approach is intended to improve detection accuracy and robustness to variation in textual data, while supporting Latvia-specific identifier formats and the linguistic characteristics of Latvian.

The methodology comprises the following stages:

1. Selection and configuration of tools;
2. Development of an algorithm for personal-data detection;
3. Adaptation of models and rules to Latvian-language texts and Latvia-specific identifiers;
4. Definition of personal-data categories;
5. Assessment of data risk and sensitivity.

3.1. Tools and Technologies Used

A set of technologies was used to develop the personal-data detection approach, which combines machine learning-based entity recognition, regular expressions, and structured-data detection methods. The approach is based on two complementary components.

The first component is an NER model based on the XLM-RoBERTa architecture and loaded using the Hugging Face Transformers library. The Davlan/xlm-roberta-base-ner-hrl model was selected because it has been trained for multiple languages, including Latvian, and provides a basis for multilingual entity recognition.

The second component is based on regular expressions and structured-data detection. Some categories of personal data have strictly defined formats, such as personal identification numbers and email addresses, which makes it possible to detect them effectively in text using deterministic methods. However, the formats of some structured data types may overlap. In such cases, the detection mechanism must be supplemented with contextual keywords and other textual indicators to determine correctly which data type a detected value represents.

3.2. Adaptation to Latvian and Latvia-Specific Identifiers

The development of this personal-data detection approach required specific adaptation to Latvian-language texts and Latvia-specific identifiers. Latvian is a morphologically rich language characterized by diacritical marks and productive word-formation patterns, which may reduce the accuracy of a general multilingual NER model unless additional processing and adaptation are applied.

3.2.1. Linguistic Adaptation

To improve the detection of entities in Latvian-language texts, several linguistic adjustments were introduced:

- Support for Latvian diacritics—all regular expressions and filters account for Latvian diacritical marks (ā, ē, ī, ū, š, ģ, ķ, ļ, ņ, ž), which supports correct matching of words containing Latvian-specific characters and improves personal-name detection.
- Morphological filters—words with typical abstract-noun suffixes (*-ums*, *-ība*, *-ācija* *-isms*) were excluded to reduce false-positive detections during personal-name detection.
- Use of the official PMLP personal-name list—the approach automatically loads the list of Latvian personal names maintained by PMLP, which supports the validation of detected given names and helps reduce false-positive classifications.
- Compound-name handling—Latvian given names and surnames may consist of multiple components or may be written with a hyphen, for example, Anna-Marija Ozoliņa. Therefore, detected name components are merged where they form a compound or hyphenated personal name.

3.2.2. Personal-Data Categories and Identifiers in Latvia

A range of personal-data categories and structured identifiers may occur in Latvian-language texts, each with its own format and context of use. Letters occurring in document identifiers are generally written in uppercase in official documents, although lowercase

forms may also occur in free text. To support accurate detection of these data categories, specialized regular expressions and context rules were developed for the relevant identifier types. The following personal-data categories and identifiers were included in the proposed approach:

- Given name and surname. Personal names may contain Latvian diacritical marks and may consist of multiple components. Given names and surnames may also contain hyphens.
- Personal identification number. A Latvian personal identification number consists of 11 digits in the format XXXXXX-XXXXX, where “X” is a number. In the older date-based format, the first six digits represent the date of birth in DDMMYY format, followed by a digit indicating the century of birth and four additional digits. They may appear in text with or without a hyphen. Both date-based formats and newer non-date-based formats must be considered.
- Email address. An email address follows a structured format consisting of a local part and a domain name and may contain Latin letters, digits, and permitted special characters.
- Telephone number. Latvian telephone numbers generally consist of eight digits and may occur with the country-code prefix “+371” or “371”.
- Address. An address may include a street name, building number, apartment number, city, parish, municipality, and postal code. Latvian postal codes follow the format LV-XXXX, where X represents a digit. Street names may contain Latvian diacritical marks.
- International Bank Account Number (IBAN). A Latvian IBAN begins with the country code LV, followed by two check digits, a four-character bank identifier, and a thirteen-digit account-number component. The full Latvian IBAN consists of 21 characters.
- Payment card number. A payment card number commonly consists of 16 digits and may be written with or without spaces.
- Card verification value (CVV). A CVV is typically a three-digit payment-card security code.
- Driver’s license number. In Latvia it commonly consists of 2 Latin letters and 6 digits.
- IP address. An IPv4 address follows the format XXX.XXX.XXX.XXX, where each component represents a numerical value from 0 to 255.
- Identity-card number or Electronic identification (eID) card number. The identity-card number consists of 2 letters followed by 7 digits.
- Passport number. The Latvian passport number consists of 2 letters followed by 7 digits.

3.3. Risk and Entropy Assessment

The assessment of text sensitivity is based on two complementary parameters:

1. Weighted risk score, which reflects the potential consequences of unauthorized disclosure of personal data and enables the sensitivity level to be assessed according to the types of personal data detected.
2. Shannon entropy, which characterizes the diversity and distribution of the detected personal-data types within the text.

This combined approach provides a quantitative assessment of the overall risk level and the diversity of the detected personal-data structure. It makes it possible not only to assess the risk associated with individual identifiers, but also to identify texts in which increased risk arises from the combination of different personal-data types rather than from the presence of a single dominant type. The methodology is informed by the principles of the GDPR and European Union guidance on data protection and provides a systematic basis for comparing the sensitivity of textual content.

3.3.1. Weighting and Risk Score Calculation

Each type of personal data is assigned a weight w_i , which reflects the level of risk associated with the presence or unauthorized disclosure of that data type. The weighting system is based on the following criteria:

- Identification potential—the extent to which the data type enables the identification of an individual;
- Potential for direct or indirect harm—the possible consequences of a personal-data breach;
- Relevance to authentication—whether the data type may be used to access resources or verify identity;
- Potential privacy impact in a DPIA context—the significance of the data type in assessing risks to the rights and freedoms of individuals.

When assigning these weights, the study adopts standard GDPR-compatible risk assessments and DPIA techniques. Such a weight allocation system considers principles set out in ISO/IEC 27002 and ISO/IEC 27701 documents, where both the identification potential of the data elements and the severity of consequences in the event of unauthorized disclosure are evaluated. As a result, high-risk identifiers receive higher weights, while lower-risk identifiers are weighed according to their more limited identification or misuse potential.

Data types assigned a high-risk weight ($w_i = 5$) are those that can directly identify an individual or may provide access to resources. These data types are particularly critical, as their unauthorized disclosure may lead to fraud, identity theft, or financial loss.

Medium-risk weights ($w_i = 3$ or $w_i = 4$) are assigned to data types that may partially identify an individual or contribute to profiling.

Low-risk weights ($w_i = 1$ or $w_i = 2$) are assigned to data types that generally present a lower risk when considered in isolation, but may become significant when combined with other data and thereby increase the overall risk level.

Table 1 summarizes the personal-data types used in this study and the risk weights assigned to them.

Table 1. Risk Weights Assigned to Personal-Data Types.

Personal-Data Type/Weight	1	2	3	4	5
Given name and surname	X				
Personal identification number					X
Email address	X				
Telephone number		X			
Address			X		
IBAN				X	
Payment card number					X
Card verification value (CVV)					X
Driver's license number				X	
IP address		X			
Identity-card number					X
Passport number					X

A total risk score is calculated for each text using the following formula:

$$R = \sum_{i=1}^n c_i \cdot w_i \quad (1)$$

where

- R is risk score;
- c_i is the number of detected instances of the i -th personal-data type;
- w_i is the risk weight assigned to the i -th personal-data type;
- n is the total number of personal-data types included in the assessment.

In this way, the risk score reflects both the number of detected personal-data instances and the assigned risk weight of each detected data type in a text. The metric enables texts to be compared, supports the identification of higher-risk cases, and helps determine which texts may require additional attention or protection measures.

Based on the calculated risk score, texts are classified into three operational sensitivity levels:

- Low sensitivity— $R < 5$;
- Medium sensitivity— $5 \leq R < 15$;
- High sensitivity— $R \geq 15$.

This classification enables the automated assignment of sensitivity levels to texts, reducing the need for manual assessment and supporting the consistent assignment of risk levels across different texts.

The sensitivity thresholds developed in this study are informed by general data-protection principles concerning risk assessment, including consideration of the severity of potential consequences and the risk of re-identification arising from specific combinations of personal data. The particular threshold values applied in this study constitute an operational model proposed for the quantitative assessment of the sensitivity of textual content.

The selected sensitivity thresholds (low, medium, high) are determined based on the distribution of the calculated risks in the assessed dataset and are aligned with practices used in conducting DPIAs, where risk categories are defined according to the likelihood and impact of potential harm. Therefore, the threshold values represent an operationalization of the GDPR risk-based approach: low sensitivity corresponds to minimal identification potential, medium sensitivity reflects a moderate risk of reidentification or misuse, and high sensitivity indicates the presence of data categories capable of causing significant harm, such as financial or national identifiers.

At the same time, these thresholds should not be regarded as universal. They may be adapted to the risk-assessment needs of a particular organization, the operational context of data processing, or the specific requirements of a particular industry. In this sense, the proposed methodology provides a structured framework that can be expanded or refined where a more stringent or more flexible assessment is required.

This approach is particularly relevant to low-resource languages and application contexts in which standardized tools and sensitivity-classification criteria are limited. The proposed system therefore provides a basis for further methodological development and may be integrated into automated data-protection workflows and manual quality-control procedures.

3.3.2. Shannon Entropy

The risk score alone does not fully characterize the composition of detected personal data, as two texts may have the same risk score while differing substantially in the distribution of detected personal-data types. Shannon entropy is therefore calculated as an additional measure to characterize the diversity and distribution of personal-data categories detected in a text.

Let:

- s_i denote the contribution of the i -th personal-data type to the total risk score, calculated as

$$s_i = c_i \cdot w_i; \quad (2)$$

- S denote the total weighted contribution of all detected personal-data types, calculated as

$$S = \sum s_i \quad (3)$$

- p_i denote the relative share of the i -th personal-data type in the total weighted contribution, calculated as

$$p_i = s_i / S \quad (4)$$

For texts in which $S > 0$, Shannon entropy is calculated as follows:

$$H = - \sum_{i=1}^n p_i \cdot \log_2 p_i \quad (5)$$

The results are interpreted as follows:

- Low entropy—one personal-data type dominates the weighted risk score; the detected risk is concentrated in a limited category of personal data.
- Medium entropy—several personal-data types contribute to the weighted risk score, although one or more types remain dominant.
- High entropy—the weighted risk score is distributed across multiple personal-data types, indicating a more diverse combination of detected personal data.

4. Implementation of the Hybrid Personal-Data Detection Approach

The developed hybrid approach was implemented in Python 3.12.10 as a modular system that combines a transformer-based NER model, regular expressions, and context filters. This architecture is intended to provide flexibility and support accurate processing of Latvian-language texts from various sources, including Microsoft Teams messages, emails, and documents. This section describes the practical implementation, code structure, and technical components of the developed system.

The NER component is based on the multilingual transformer model Davlan/xlm-roberta-base-ner-hrl, loaded using the Hugging Face Transformers library. The model was used without additional fine-tuning, as no sufficiently large manually annotated Latvian-language NER corpus was available for this study. Instead, the model output was supplemented with language-specific post-processing rules, including token merging, morphological filtering, and validation against the official PMLP list of Latvian personal names.

Prior to inference, all texts were normalized: non-breaking whitespace characters were replaced, redundant whitespace was removed, and Latvian diacritics were preserved. As shown in Figure 1, the model was executed using the token-classification pipeline with the parameter `aggregation_strategy = "simple"`, which aggregates subword token predictions into complete entity spans. Inference was performed on a CPU-based system equipped with an Intel i7 processor and 32 GB of RAM. Processing time ranged from approximately 20 to 40 ms per text fragment, depending on text length.

```
self.ner = pipeline(
    "token-classification",
    model=self.model,
    tokenizer=self.tokenizer,
    aggregation_strategy="simple",
    device=device
)
```

Figure 1. NER pipeline configuration (code snippet).

4.1. Architecture of the Hybrid Personal-Data Detection Approach

The implemented architecture is built around the *LatvianPIIDetector* class, which provides a consistent text-processing flow and integrates the main detection components into a single structure (Figure 2).

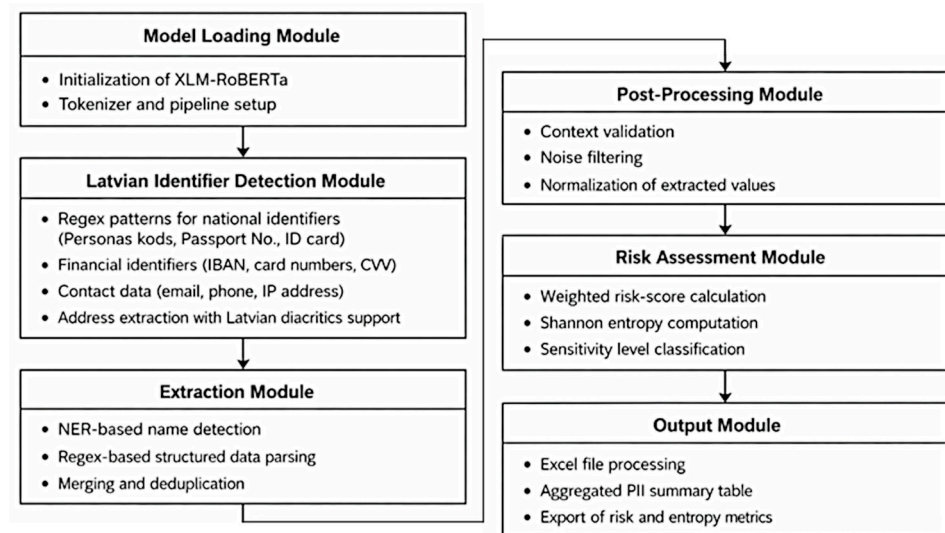


Figure 2. Architecture of the Hybrid Personal-Data Detection Approach.

4.2. Implementation of Personal-Data Detectors in Python

The practical implementation of the hybrid approach is based on a clearly defined set of detectors, with each personal-data type detected using a specific mechanism implemented in Python: a NER model, regular expressions, or context filters. This structured approach provides a transparent and repeatable detection process and facilitates the extension of the system by adding new data types.

To ensure consistent detector management and processing flow, all detection mechanisms are integrated into the *LatvianPIIDetector* class, where each data type is associated with a corresponding detection mechanism.

Each detector is implemented either as a separate element in the regular-expression dictionary or as a function responsible for detecting a specific entity type. This approach provides precise control over how each personal-data type is detected and helps reduce false positives through the use of additional context filters. In addition, this structure supports system modularity, since each detector can be updated or improved independently of the other components.

The detection of given names and surnames was performed by the NER component, supplemented with several additional constraints. To reduce false-positive personal-name detections, the system automatically loads the official list of Latvian personal names maintained by the Citizenship and Migration Affairs Office (PMLP) (Figure 3). This enables the system to verify whether a given name detected by the NER model occurs in the official Latvian personal-name list, thereby helping to reduce false positives.

To reduce false-positive cases in which the NER model incorrectly classifies common nouns as given names, a morphological filtering mechanism was developed based on an analysis of Latvian suffix patterns. Several Latvian suffixes, such as *-ums*, *-ība*, *-ācija*, and *-isms*, are characteristic of abstract nouns and rarely occur in given names (Figure 4). Consequently, words ending in these suffixes are excluded from the set of potential given names, thereby helping to reduce false-positive detections.

```
@lru_cache(maxsize=1)
def load_latvian_first_names(self):
    url = "https://www.pmlp.gov.lv/lv/personvardu-datu-baze"
    try:
        r = requests.get(url, timeout=10)
        r.raise_for_status()
    except Exception:
        return set()
    html = r.text
    names = set(re.findall(r">([A-ZĀČĒĢĪĶĻŅŌŖŠŪŽ][a-zāčēģīķļņōŗšūž])+<",
    html))
    return names
```

Figure 3. Implementation of the PMLP Personal-Name List Loader.

```
if word.lower().endswith(("ums", "iba", "isms", "ācija")):
    return False
```

Figure 4. Latvian Suffix-Based Morphological Filter.

Additional validation is performed by checking whether each potential given name satisfies several orthographic and formal criteria. At this stage, the system checks whether the word begins with an uppercase letter, contains no digits or non-alphabetic characters, and falls within the length range defined for Latvian given-name candidates (Figure 5). These criteria help exclude cases in which the NER model would otherwise incorrectly classify common nouns as given names.

```
if not re.match(r"^[A-ZĀČĒĢĪĶĻŅŌŖŠŪŽ][a-zāčēģīķļņōŗšūž]+$", word):
    return False
if len(word) < 3 or len(word) > 25:
    return False
```

Figure 5. Orthographic and Length Validation of Potential Given Names.

The structure of Latvian addresses is complex and differs from address formats for which many general-purpose detection tools are designed. A Latvian address may contain several components, including a street name, street type, building number, apartment number, city or municipality, and postal code. Street types may occur in forms such as *ielā*, *gatve*, *bulvāris*, and *prospekts*. In addition, street names may contain Latvian diacritical marks, compound forms, and inflectional endings, all of which complicate automatic processing and limit the applicability of rules not specifically adapted to Latvian address formats.

To support accurate address detection, a language-specific regular expression was developed to capture common Latvian address patterns and orthographic characteristics (Figure 6). This expression accounts for common street-type forms, building and apartment number formats, and the structure of Latvian postal codes. It therefore provides a Latvian-specific mechanism for detecting complete address expressions rather than relying on general language-agnostic patterns.

```
self.address_full_regex = re.compile(
    r"\b([A-ZĀČĒĢĪĶĻŅŌŖŠŪŽa-zāčēģīķļņōŗšūž]+"
    r"(?:\s+(?:ielā|ielā|gatve|prospekts|bulvāris|bulvārī|linija|ceļš))"
    r"\s+\d+[A-Za-z](?:[-/]\d+)?"
    r"(?:,\s*[A-ZĀČĒĢĪĶĻŅŌŖŠŪŽa-zāčēģīķļņōŗšūž])*"
    r"(?:,\s*LV-\d{4})?\b"
)
```

Figure 6. Regular Expression for Full Latvian Address Detection.

5. Evaluation and Results of the Hybrid Personal-Data Detection Approach

To evaluate the effectiveness of the hybrid approach, an experimental evaluation was conducted using a dataset created within the project and based on authentic organizational data from Microsoft Teams messages, emails, and documents. Since the initial dataset contained a limited number of personal-data instances, it was supplemented with synthetic texts containing various categories of Latvian personal data. The dataset included texts containing given names and surnames, personal identification numbers, telephone numbers, email addresses, addresses, International Bank Account Numbers (IBANs), payment card numbers and card verification values (CVVs), IP addresses, identity-card numbers, passport numbers, and driver's license numbers. Particular attention was paid to detectors whose matching patterns may overlap, for example, when distinguishing between passport numbers and identity-card numbers. In such cases, additional validation and contextual analysis were applied to improve classification accuracy and reduce false positives.

The dataset contains a total of 13,525 records, of which 12,325 were derived from authentic organizational data and 1200 were synthetically generated. All text samples were stored in an Excel file and manually annotated by identifying personal-data instances and assigning them to the corresponding categories. The annotation was performed by two researchers with expertise in information security and Latvian-language processing to support the consistency and reliability of the labeling process.

The aim of the experiment was to assess the detection performance of the hybrid approach in unstructured texts and to evaluate the robustness of the detectors under format variation, orthographic variation, and differences in context.

After execution of the implemented system, an Excel-formatted output file was generated, indicating which types of personal data were detected in each record and how many instances of each type were identified. This output file was then compared with the manually annotated reference dataset to evaluate the performance of the detectors. The annotators manually reviewed the automatically generated results against the reference annotations and, on the basis of this comparison, compiled the detailed results. This manual validation provides a basis for the subsequent analysis of the performance of the hybrid approach.

Table 2 illustrates how the hybrid approach assesses the sensitivity of different texts on the basis of the number and diversity of personal-data instances detected in them. It includes examples from the larger dataset on which the developed approach was evaluated and indicates the risk score, entropy, and operational sensitivity level calculated for each text. The examples illustrate the relationship between the weighted quantity and diversity of detected personal data and the assigned risk level. Texts containing only a small number of detected personal-data instances, such as a personal name, an email address, or a single structured identifier, may be assigned a low operational sensitivity level. In such cases, the risk score is low, and the detected data types are relatively homogeneous, which is reflected in a low entropy value. Texts containing multiple higher-risk identifiers, such as a personal identification number, passport number or identity-card number, address, and contact information, may achieve a higher risk score and higher entropy. Such combinations may be classified as medium- or high-sensitivity texts.

In the specific example in the first record, the calculated risk score is 2, despite the presence of an identity-card number in the text. This outcome occurs because the hybrid approach did not recognize the ID-card number in this instance and therefore did not include it in the weighted risk calculation. As a result, the system detected only two low-risk identifiers: a personal name and an email address, which led to a low overall risk score and a correspondingly low operational sensitivity level.

Table 2. Risk and Sensitivity Analysis of Personal Data in Text Examples.

Text Example	Risk Score	Entropy	Operational Sensitivity Level
“Labdien! Mani sauc Mārtiņš Eglītis. Reģistrējoties jūsu portālā, nepieciešams apstiprināt identitāti. Mans ID kartes numurs ir LA1234567. Dzimšanas datums: 14.02.1990 E-pasts: martins.eglitis9@gmail.com Ar cieņu, Mārtiņš Eglītis”	2	1.00	Low
“Labdien! Reģistrējoties jūsu portālā, nepieciešams apstiprināt identitāti. Esmu Mārtiņš Eglītis, dzimis 1990. gada 14. februārī, un mans ID kartes numurs ir LA1234567. Dzīvoju Rīgā, un ar mani var sazināties pa e-pastu martins.eglitis90@gmail.com. Lūdzu, informējiet, ja nepieciešami papildu dokumenti. Ar cieņu, Mārtiņš Eglītis”	5	0.72	Medium
“Labdien! Esmu Rūdolfs Vīksne, dzimis 13.03.1983. (pers. kods 830313-14700), ID nr. LV55662233, dzīvoju Tērbatas ielā 18-2, Rīgā. Ar nožēlu paziņoju, ka tikšanās 27.07.2025. plkst. 11:00 jūsu birojā jāatceļ slimības dēļ. Lūdzu piedāvāt jaunu tikšanās laiku augustā. Par termiņa maiņu varat sazināties pa +371 26455002 vai rudolf.viksne@inbox.lv. Ar cieņu, Rūdolfs Vīksne”	14	2.16	Medium
“Labdien! Esmu Martins Kalniņš, dzimis 15.09.1982. (pers. kods 820915-14568), pase nr. LV22330011, dzīvoju Skolas ielā 3-2, Sigulda. Interesē optiskā interneta pieslēgums 300 Mb/s privātmājā. Lūdzu nosūtīt tehniskās prasības un cenu piedāvājumu. Depozītu 30 EUR lūdzu pārskaitīt uz kontu LV63HABA0551020304301. Par detaļām—+371 26455021 vai martins.kalnins@inbox.lv. Paldies, Martins Kalniņš”	18	2.44	High

Table 2 illustrates how the hybrid approach can be used not only to detect personal data, but also to quantitatively assess the operational sensitivity level of textual content by considering both the number of detected instances and the combination of detected personal-data types. This enables the automatic distinction between texts assigned a low operational sensitivity level and those that may require additional data-protection measures.

The detector-level evaluation results are summarized in Table 3, which reports the number of true positives (TPs), false positives (FPs), and false negatives (FNs) for each personal-data type. To provide a more interpretable assessment of detector performance,

precision, recall, and F1-score were calculated for each category. Precision indicates the proportion of detected instances that were classified correctly, recall indicates the proportion of reference instances that were successfully detected, and F1-score provides a combined measure of precision and recall. These indicators enable a comparative analysis of detector performance and provide an empirical basis for assessing the strengths and weaknesses of the developed approach, as well as for identifying areas for future improvement.

Table 3. Per-Category Personal-Data Detection Results.

Personal-Data Type	TP	FP	FN	Number of Reference Instances	Precision (%)	Recall (%)	F1-Score (%)
Given name and surname	16,787	4276	1800	18,587	79.70	90.32	84.68
Personal identification number	14,945	6	44	14,989	99.96	99.71	99.83
Email address	15,494	22	154	15,648	99.86	99.02	99.44
Telephone number	14,618	74	13	14,631	99.50	99.91	99.70
Address	12,612	6	122	12,734	99.95	99.04	99.50
IBAN	12,216	11	2	12,218	99.91	99.98	99.95
Payment card number	6	2	2	8	75.00	75.00	75.00
Card verification value (CVV)	2	4	0	2	33.33	100.00	50.00
Driver's license number	7	4465	0	7	0.16	100.00	0.31
IP address	39	39	0	39	50.00	100.00	66.67
Identity-card number	11	0	151	162	100.00	6.79	12.72
Passport number	7420	0	4627	12,047	100.00	61.59	76.23

Table 3 also provides insight into the error structure of the developed approach. High numbers of false positives or false negatives for particular personal-data types may indicate overlap between identifier formats, insufficient contextual rules, or variations in the representation of those data types. These results therefore help identify the detectors that require further refinement.

The results summarized in Table 3 indicate that the hybrid approach performs strongly for several personal-data types with well-defined structured formats. Personal identification numbers achieved an F1-score of 99.83%, email addresses 99.44%, telephone numbers 99.70%, addresses 99.50%, and IBANs 99.95%. These results indicate that regular-expression-based detection is highly effective for categories characterized by clearly defined formats and limited structural variation.

The detection of given names and surnames also produced comparatively strong results, with a precision of 79.70%, a recall of 90.32%, and an F1-score of 84.68%. The lower precision indicates that, although most reference instances were detected, a substantial number of other text elements were incorrectly classified as personal names. This finding supports the need for further refinement of contextual filters and language-specific validation mechanisms.

Substantially weaker performance was observed for document identifiers with overlapping formats. The driver's license number detector achieved a recall of 100.00%, but its precision was only 0.16%, resulting in an F1-score of 0.31%. This indicates that, although all driver's license numbers present in the reference dataset were detected, the detector incorrectly classified a very large number of other identifiers as driver's license numbers. The passport number detector achieved a precision of 100.00% but a recall of only 61.59%, while the identity-card number detector achieved a precision of 100.00% but a recall of only 6.79%. These results indicate that the current rules for distinguishing among document

identifiers require substantial improvement, particularly through refined conflict-resolution logic and contextual validation.

The results for payment card numbers, card verification values, and IP addresses should be interpreted cautiously because these categories are represented by relatively small numbers of reference instances. Although their calculated metrics provide an initial indication of detector behavior, additional evaluation data would be required to draw reliable conclusions about their general performance.

During system development, it was anticipated that identity-card numbers and passport numbers might have overlapping formats, creating a risk that the same identifier would be classified under both categories. To address this issue, the implementation was supplemented with an additional verification mechanism intended to prevent the same document number from being recorded simultaneously in both categories (Figure 7).

```
if "Pases Nr" in results and "ID kartes Nr" in results:
    passport_nums = set(results["Pases Nr"])
    results["ID kartes Nr"] = [
        x for x in results["ID kartes Nr"] if x not in passport_nums
    ]
```

Figure 7. Deduplication of Passport and Identity-Card Numbers.

The high number of false positives for driver's license numbers can be explained by insufficient differentiation between document-identifier formats in the implemented detection rules. In particular, the evaluation results suggest that a substantial number of passport numbers were incorrectly classified as driver's license numbers. Although passport numbers and driver's license numbers are intended to follow distinct structural patterns, their similar alphanumeric structure and identical overall character length can create ambiguity when broad regular-expression rules are applied without sufficient contextual validation.

Although the system includes an overlap-filtering mechanism for the passport number ("Pases Nr." in Latvian) and driver's license number ("Vadītāja apliecības Nr." in Latvian) categories, this mechanism functions only when both detectors return the same identifier in an identical textual representation (Figure 8). In practice, differences in formatting, spacing, or pattern matching may prevent the filter from identifying equivalent values. Consequently, a value corresponding to a passport number may remain classified as a driver's license number, contributing to the high number of false positives observed for the latter category and the high number of false negatives observed for passport numbers.

```
if "Vadītāja apliecības Nr." in results and "Pases Nr." in results:
    dl_nums = set(results["Vadītāja apliecības Nr."])
    results["Pases Nr."] = [
        x for x in results["Pases Nr."] if x not in dl_nums
    ]
```

Figure 8. Overlap Filtering for Passport Number and Driver's License Number Detections.

These results indicate that regular-expression-based detection alone is insufficient for reliably distinguishing between document identifiers with similar formats. Further improvement requires normalization of detected values before overlap checking, refinement of conflict-resolution rules, and the use of additional contextual indicators to distinguish between passport numbers and driver's license numbers.

It is also important to note that a text does not always contain a keyword that enables the system to determine unambiguously the personal-data category to which a detected string belongs. As a result, a regular-expression detector may successfully extract a data sequence but may not classify it correctly in the absence of additional semantic cues. This

limitation highlights the need for contextual information to support reliable differentiation between personal-data types with similar formats.

Comparative Analysis of the Proposed Hybrid Approach and Microsoft Presidio

To compare the developed hybrid approach with Microsoft Presidio, two representative Latvian text examples were selected from the constructed dataset (Figure 9). These examples illustrate the behavior of the systems on texts containing different combinations of personal-data types and provide an illustrative comparison of the strengths and limitations of the developed approach.

1) Labdien, Artur,

Jūsu adrese dokumentos: Liepāja, Peldu iela 8-12, LV-3401.

Personas kods: 150393-56789

Ja nepieciešams mainīt informāciju, lūdzu, sazinieties: info@bank.lv.

2) Labdien, Laura,

Mēs nosūtām Jums līguma projektu.

Personas kods: 220685-12345

Līgums tiks parakstīts attālināti, tāpēc lūdzu, pārliecinieties, ka Jūsu e-pasts laura.k@example.lv ir aktīvs un tālruņa numurs ir pieejams: +371 27775544.

Figure 9. Representative Text Examples Used for Comparison.

Both examples were processed using Microsoft Presidio and the developed personal-data detection system adapted specifically to Latvian-language texts. This comparison illustrates how the two systems detect personal data in selected communication scenarios and highlights limitations that may arise when a general-purpose tool is applied without Latvia-specific recognizers and contextual rules.

Microsoft Presidio provides functionality for detecting and anonymizing personally identifiable information and can be extended with additional recognizers and language-specific components. However, its default configuration contains recognizers and models for English, while support for additional languages requires adaptation of the NLP engine and the relevant recognizers. In the evaluated configuration, Presidio did not include rules specifically designed for Latvian morphology, Latvian address forms, or Latvia-specific document identifiers. Latvian personal names may contain diacritical marks and inflected forms, street names and street-type terms are language-specific, and document numbers follow national formats. These characteristics may reduce detection performance when Latvia-specific recognizers and context rules are not configured.

To ensure comparability, the two systems were compared under the same circumstances—the same input data, the same annotation procedures, and the same evaluation criteria. In this way, any differences in performance will result from differences in the underlying systems, not the evaluation framework.

Furthermore, Latvian is a low-resource language for which the availability of large, high-quality annotated datasets for training or adapting NER models is limited. This creates additional challenges for detecting personal data in Latvian-language texts. In the present comparison, the observed results should therefore be interpreted as a comparison between the developed Latvian-specific approach and the evaluated Presidio configuration, rather than as a general assessment of the capabilities of Microsoft Presidio.

As illustrated in Figures 10 and 11, the evaluated Presidio configuration correctly detected the given names, surnames, and email addresses present in the two text examples. It also detected the telephone numbers in these examples, although it additionally classified

some unrelated digit sequences as telephone numbers. However, the Latvian personal identification number was incorrectly classified as a telephone number. This constitutes a significant classification error and indicates that the evaluated configuration did not include appropriate support for this Latvia-specific identifier. This form of classification error would be highly relevant in a GDPR-compliant scenario where personal ID card numbers are considered sensitive identifiers. The incorrect risk assessment may occur in this scenario, leading to either a wrong risk score or under-anonymized data. In the case of addresses, Presidio detected individual components separately but did not combine them into a complete address expression and did not detect the building and apartment numbers. In addition, the evaluated configuration did not recognize Latvian street-type terms, such as *ielā*, *bulvāris*, *prospekts*, and *gatve*, whose forms and inflections differ from those used in English-language address patterns.

Input

Enter text

Labdien, Artur,
Jūsu adrese dokumentos: Liepāja, Peldu iela 8-12, LV-3401.
Personas kods: 150393-56789
Ja nepieciešams mainīt informāciju, lūdzu, sazinieties: info@bank.lv.

Output

De-identified

Labdien, <PERSON>,
Jūsu adrese dokumentos: <LOCATION>, <LOCATION> iela 8-12, <LOCATION>.
Personas kods: <PHONE_NUMBER>
Ja nepieciešams mainīt informāciju, lūdzu, sazinieties: <EMAIL_ADDRESS>.

Findings

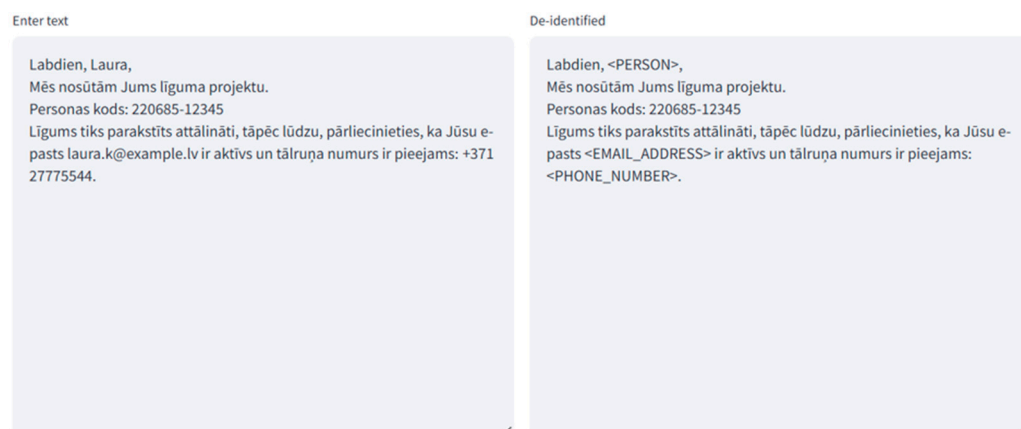
	Entity type	Text	Start	End	Confidence
0	LOCATION	Liepāja	40	47	1
1	LOCATION	LV-3401	66	73	1
2	EMAIL_ADDRESS	info@bank.lv	159	171	1
3	ORGANIZATION	bank.lv	164	171	1
4	PERSON	Artur	9	14	0.99
5	LOCATION	Peldu	49	54	0.99
6	URL	bank.lv	164	171	0.5
7	PHONE_NUMBER	150393-56789	90	102	0.4

Figure 10. Microsoft Presidio Results for the First Text Example.

These results indicate that the evaluated Presidio configuration was not specifically adapted to Latvian address structures or Latvia-specific identifier formats. The outputs shown in the figures indicate that Presidio correctly detected several common types of personal data in the selected examples, including given names and surnames, email addresses, and telephone numbers. However, the Latvian personal identification number was not classified correctly, and in some cases the tool classified domain names as URLs or organization names while assigning low confidence scores to some structured identifiers. These observations highlight a limitation of the evaluated configuration: without Latvian-specific recognizers and contextual rules, the system did not reliably distinguish several local identifiers and address components in the selected examples. This mismatch between the configured recognizers and the linguistic and structural characteristics of Latvian data resulted in classification errors and reduced detection performance in the evaluated cases.

In the same examples, the developed approach correctly detected the personal-data types included in the comparison (Figure 12). In particular, the system detected complete address expressions rather than isolated address components, and it correctly classified the Latvian personal identification numbers present in the examples. These results illustrate the benefit of applying Latvian-specific regular-expression rules, morphological filters, and post-processing mechanisms when detecting personal data in Latvian-language

texts. In the selected examples, the developed approach detected Latvian address expressions and Latvia-specific identifiers more effectively than the evaluated Microsoft Presidio configuration.



Findings

	Entity type	Text	Start	End	Confidence
0	EMAIL_ADDRESS	laura.k@example.lv	159	177	1
1	PERSON	Laura	9	14	0.99
2	URL	example.lv	167	177	0.5
3	PHONE_NUMBER	+371 27775544	219	232	0.4

Figure 11. Microsoft Presidio Results for the Second Text Example.

source_row	Name_count	Name	Person_ID_count	Person_ID	Email_count	Email	Phone_count	Phone	Address_count	Address
1	1	Artur	1	150393-56789	1	info@bank.lv	0		1	Liepāja, Peldu iela 8-12, LV-3401
2	1	Laura	1	220685-12345	1	laura.k@example.lv	1	371 27775544	0	

Figure 12. Results of the Developed Hybrid Approach for Both Text Examples.

The performance of both detection methods is summarized in Table 4, which provides a structured, side-by-side comparison across the main personal-data categories present in the two representative examples. In addition to correctly detected entities, the table also documents systematic errors observed in the Microsoft Presidio configuration, including the misclassification of Latvia-specific identifiers and fragmented address detection. By contrast, the developed hybrid approach successfully identified complete address expressions and correctly classified national identifiers.

Table 4. Summary of Detection Performance in the Two Representative Examples.

Data Type	Example	Presidio	Developed Hybrid Approach
Personal names	Laura; Artur	Correct	Correct
Email addresses	info@bank.lv; laura.k@example.lv	Correct	Correct
Phone numbers	37127775544	Correct, but +1 false positive (150393-56789)	Correct
Personal identification number	150393-56789; 220685-12345	Unsupported identifier type	Correct
Address	Liepāja, Peldu iela 8-12, LV-3401	Fragmented, partial (Liepāja, Peldu, LV-3401)	Complete, merged

6. Conclusions

In this work, a hybrid approach for the automatic detection of personal data in Latvian-language texts was developed and experimentally evaluated. The approach combines a transformer-based NER model with regular expressions, specialized dictionaries, and context filters. The results indicate that this combination is suitable for detecting different categories of personal data in Latvian-language texts, particularly in a context where language-specific resources and detectors for Latvia-specific identifiers are limited.

The evaluation results demonstrate strong performance for several structured-identifier types. The developed approach achieved F1-scores of 99.83% for personal identification numbers, 99.44% for email addresses, 99.70% for telephone numbers, 99.50% for addresses, and 99.95% for IBANs. These findings indicate that deterministic detection methods are effective for personal-data categories with clearly defined structural formats. The detection of given names and surnames achieved an F1-score of 84.68%, indicating good overall performance but also revealing the continued presence of false-positive detections.

The risk-score and entropy-based assessment method introduced in this study adds an additional dimension to personal-data analysis. It enables texts to be assigned operational sensitivity levels based on both the weighted number of detected personal-data instances and the diversity of the detected data types. This method provides a structured basis for comparing textual content according to its potential data-protection risk and may support risk-assessment and security-policy workflows.

The contribution of this work lies in the development of a practical personal-data detection approach for Latvian-language texts and in demonstrating how NER-based processing can be combined with rule-based detectors for Latvia-specific structured identifiers. The experimental results suggest that this combination is useful for addressing detection tasks that cannot be handled uniformly by either contextual entity detection or structured-pattern matching alone. This is particularly relevant for low-resource languages, which are characterized by limited annotated data and a lack of specialized NLP tools.

At the same time, the evaluation identified substantial limitations in the detection of document identifiers with overlapping formats. The driver's license number detector achieved an F1-score of only 0.31% because of an extremely high number of false positives, while identity card numbers achieved an F1-score of 12.72% because most reference instances were not detected. Passport numbers achieved an F1-score of 76.23%, with reduced recall indicating that a considerable number of instances were missed. These findings show that additional contextual validation and conflict-resolution mechanisms are required to distinguish reliably among similar document-identifier formats.

Overall, the results indicate that a locally adapted hybrid approach can support the detection of personal data in Latvian-language digital content, particularly for structured identifiers and Latvian-specific address patterns. The study also demonstrates that effective personal-data detection in a low-resource language requires category-specific evaluation, since strong aggregate results may coexist with substantial weaknesses in individual detectors. The developed approach therefore provides a basis for further refinement of language-specific personal-data detection technologies and their application in organizational data-protection workflows.

Author Contributions: Conceptualization, J.S. and H.G.; methodology, J.S.; software, J.S. and H.G.; validation, J.S., H.G., J.K. and V.Z.; formal analysis, J.S. and H.G.; investigation, J.S.; resources, J.K. and V.Z.; data curation, J.S., H.G., J.K. and V.Z.; writing—original draft preparation, J.S. and H.G.; writing—review and editing, J.S. and H.G.; visualization, H.G.; supervision, A.R.; project administration, A.R. and J.K.; funding acquisition, A.R. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project “Competence Centre of Information and Communication Technologies for Digitalization” of the Recovery and Resilience Facility, contract No. 2.2.1.3.i.0/1/24/A/CFLA/006 signed between IT Competence Centre and Central Finance and Contracts Agency, Research No. 1.1 “Analysis, Evaluation, and Classification of User-Generated Digital Content in IT Systems Using AI/ML Approaches”.

Data Availability Statement: The datasets presented in this article are not publicly available because they include real organizational communications and documents that may contain personal data, confidential business information, and commercially sensitive content. Due to privacy, data-protection, and commercial confidentiality restrictions, the dataset cannot be shared in full. Requests for access to aggregated or anonymized information supporting the findings of this study may be directed to the corresponding author.

Conflicts of Interest: Authors Jurijs Kornijenko and Vitālijs Zabiņako were employed by the company SIA “ABC software”. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Official Journal of the European Union. Regulations. 27 April 2016. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679> (accessed on 22 April 2026).
2. Cookie Script. GDPR Data Protection Impact Assessment (DPIA). Available online: <https://cookie-script.com/blog/gdpr-data-protection-impact-assessment-dpia> (accessed on 22 April 2026).
3. State of California Department of Justice; Bonta, R. California Consumer Privacy Act (CCPA). Available online: <https://oag.ca.gov/privacy/ccpa> (accessed on 22 April 2026).
4. Sistema Cofeci Creci. LGPD—Lei Geral de Proteção de Dados. Available online: <https://www.cofeci.gov.br/lgpd> (accessed on 22 April 2026).
5. ISO/IEC 27001:2022 (en); Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements. ISO: Vernier, Switzerland, 2022. Available online: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27001:ed-3:v1:en> (accessed on 22 April 2026).
6. ISO/IEC 27002:2022 (en); Information Security, Cybersecurity and Privacy Protection—Information Security Controls. ISO: Vernier, Switzerland, 2022. Available online: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27002:ed-3:v2:en> (accessed on 22 April 2026).
7. ISO/IEC 27701:2025 (en); Information Security, Cybersecurity and Privacy Protection—Privacy Information Management Systems—Requirements and Guidance. ISO: Vernier, Switzerland, 2025. Available online: <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27701:ed-2:v1:en> (accessed on 22 April 2026).
8. Supriyadia, D. The Regulation of Personal and Non-Personal Data in the Context of Big Data. *J. Hum. Rights Cult. Leg. Syst.* **2023**, *3*, 33–69. [CrossRef]
9. Alhefeiti, A.; Aldahmani, S.; Alyammahi, A.; Alzeyoudi, S.; Alraeesi, A. Civil Protection of Personal Electronic Data in the Emirati Legislation. *Res. J. Adv. Humanit.* **2025**, *6*. [CrossRef]
10. Moric, Z.; Dakic, V.; Djekic, D.; Regvart, D. Protection of Personal Data in the Context of E-Commerce. *J. Cybersecur. Priv.* **2024**, *4*, 731–761. [CrossRef]
11. Utami, T.K.; Suryanto, S.O.; Putri, K.A.; Asriani, F. Personal Data Breach Cases in Indonesia: Perspective of Personal Data Protection Law. *Cust. Law J.* **2025**, *2*, 21. [CrossRef]
12. Diamantopoulou, V.; Tsohou, A.; Karyda, M. From ISO/IEC 27002:2013 Information Security Controls to Personal Data Protection Controls: Guidelines for GDPR Compliance. In *Computer Security*; Springer: Berlin/Heidelberg, Germany, 2020. [CrossRef]
13. Viksna, R.; Skadiņa, I. Multilingual Transformers for Named Entity Recognition. *Balt. J. Mod. Comput.* **2022**, *10*, 457–469. [CrossRef]
14. Erol, S.E.; Sağiroğlu, S.; Demirezen, M.U. Detecting Personal Health Data Disclosures in Turkish Social Data. *Int. J. Inf. Secur. Sci.* **2022**, *11*, 69–84. Available online: <https://izlik.org/JA75XB44DL> (accessed on 22 April 2026). [CrossRef]
15. Akbasli, I.T.; Beck, K.L.; Liou, W.; Tandon, A. Hybrid Rule-Based and On-Premises LLM Pipeline for Extracting CMR and CPET Metrics from Free-Text Reports in Repaired Tetralogy of Fallot. *medRxiv* **2026**. [CrossRef] [PubMed]
16. Coldea, A.; Elmessary, M.A.; Thayer, D.S.; Scanlon, I.; Davies, H. Sensitive Data Flagging within Data Quality Reports: R and Regex Integration for Effective Text Flagging in Large Datasets. *Int. J. Popul. Data Sci.* **2024**, *9*. [CrossRef]

17. Krishnaram, P.C.M. *Regex in Action: Practical Examples for Python Programmers*. 2025. Available online: https://www.researchgate.net/publication/395268363_Regex_in_Action_Practical_Examples_for_Python_Programmers (accessed on 22 April 2026).
18. Tahir, H.; Brezillon, P. Context-Based Personal Data Discovery for Anonymization. In *Context-Based Personal Data Discovery for Anonymization*; John Wiley & Sons: Hoboken, NJ, USA, 2025. [CrossRef]
19. Microsoft. Microsoft Purview. Available online: <https://learn.microsoft.com/en-us/purview/> (accessed on 22 April 2026).
20. Microsoft. Microsoft Presidio. Available online: <https://microsoft.github.io/presidio/> (accessed on 22 April 2026).
21. Google Cloud. Sensitive Data Protection Documentation. Available online: <https://docs.cloud.google.com/sensitive-data-protection/docs> (accessed on 22 April 2026).
22. AWS. Amazon Macie User Guide: Discovering Sensitive Data with Macie. Available online: <https://docs.aws.amazon.com/macie/latest/user/data-classification.html> (accessed on 22 April 2026).
23. IBM. IBM Guardium Data Protection. Available online: <https://www.ibm.com/products/guardium-data-protection> (accessed on 22 April 2026).
24. Collibra. Available online: <https://www.collibra.com/> (accessed on 22 April 2026).
25. Malekela, M.; Simon, A. Data Protection in Employment: Implications of the Personal Data Protection Act in Tanzania. *Labour Law. Issues* **2025**, *11*, C.1–C.28. [CrossRef] [PubMed]
26. Arifin, D.; Widiarty, W.S.; Nainggolan, R.M. Legal Protection of Consumer Personal Data in Business in the Digital Era. *J. Soc. Res.* **2025**, *4*, 2024–2039. [CrossRef]
27. Sulthanah, L.T.; Rasji, R. The Urgency of Establishing a Personal Data Protection Agency in Indonesia. *Law Dev. J.* **2025**, *7*, 675–687. [CrossRef]
28. Patel, D.; Johnson, A.P. Detecting Narcissistic Personality Disorder (NPD): A Hybrid Regex and NLP-Based AI Approach with Phase-Aware Classification. *IEEE Access* **2025**, *13*, 185011–185024. [CrossRef]
29. Subramani, N.; Ghate, K.; Diab, M. Personal Information Parroting in Language Models. *arXiv* **2026**, arXiv:2602.20580. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.