**International Journal of Information Systems and Project Management (IJISPM)**
2025, 13(2), e2, DOI: 10.12821/ijispm130202

RESEARCH ARTICLE

# Factors related to GDPR compliance promises in privacy policies: A machine learning and NLP approach

**Abdel-Jaouad Aberkane**
Ghent University, Tweekerkenstraat 2, 9000, Ghent, Belgium, abdeljaouad.aberkane@ugent.be

**Seppe vanden Broucke**
Ghent University, Tweekerkenstraat 2, 9000, Ghent, Belgium, seppe.vandenbroucke@ugent.be

**Geert Poels**
Ghent University, Tweekerkenstraat 2, 9000, Ghent, Belgium, geert.poels@ugent.be

**Abstract**
This paper employs Machine Learning (ML) and Natural Language Processing (NLP) techniques to examine the relationship between organizational factors, such as company size and headquarters location, of data processing entities and their GDPR compliance promises as disclosed in privacy policies. Our methodology comprises three main stages, each representing a key contribution. Firstly, we developed five NLP-based classification models with precision scores of at least 0.908 to assess different GDPR compliance promises in privacy policies. Secondly, we have collected a data set of 8,614 organizations in the European Union containing organizational information and the GDPR compliance promises derived from the organization's privacy policy. Lastly, we have analyzed the organizational factors correlating to these GDPR compliance promises. The findings reveal, among other things, that small or medium-sized enterprises negatively correlate with the disclosure of two GDPR privacy policy core requirements. Moreover, as a headquarters location, Denmark performs best regarding positively correlating with disclosing GDPR privacy policy core requirements, whereas Spain, Italy, and Slovenia negatively correlate with multiple requirements. This study contributes to the novel field of GDPR compliance, offering valuable insights for policymakers and practitioners to enhance data protection practices and mitigate non-compliance risks.

**Keywords**
general data protection regulation; privacy; privacy policy; natural language processing; machine learning.

Received: 13 February 2024 | Accepted: 23 October 2024

## 1. Introduction

The landscape of data protection has experienced notable changes since 2016—the year in which the European Union (EU) decreed the General Data Protection Regulation (GDPR)—as data processing entities processing data of EU data subjects were obliged to meet the requirements of the GDPR (Tikkinen-Piri et al., 2018). This transformation stimulated a wide range of responses, ranging across the exploration of innovative solutions like blockchain for GDPR-aligned data repositories (Al-Abdullah et al., 2020), investigating how information security frameworks can assist in implementing GDPR in the banking industry (Serrado et al., 2020), to data processing organizations reevaluating and enhancing their privacy policies (Amos et al., 2021). The latter resulted in increased length and content of the privacy policies, including the disclosure of GDPR-relevant requirements (Linden et al., 2020). This development is also reflected by the use of GDPR terminology in privacy policies, demonstrating the imprint of the GDPR on online privacy (Degeling et al., 2019; Strzelecki & Rizun, 2020).

Privacy policies act as a conduit for the data processing practices of the corresponding organizations and are, therefore, the essential source of information for the user to learn what happens to user data (Reidenberg et al., 2015) and whether it is treated in line with the regulations of the GDPR. To ensure the latter, organizations with an online presence added new privacy policies to their website, leading to an uprising in transparency on the web. In some European countries, for example, 15.7% of websites added new privacy policies close to May 25, 2018—the date of entry into force of the GDPR—whereas over 70% of websites updated their existing privacy policies (Degeling et al., 2019).

However, the surge in transparency was not accompanied by a similar rise, content-wise: only a modest portion (32%) of privacy policies examined post-GDPR fully meet its mandates (Rahat et al., 2022). A preliminary study by Contissa et al. (2018) even suggests that none of the analyzed privacy policies were fully compliant. This non-compliance issue can be related to several organizational factors, e.g., size and resources (Texeira, 2019; Freitas & Da Silva, 2018), sector, and geographical location (Zaeem & Barber, 2021). However, to our knowledge, no study analyzes the underlying organizational factors related to GDPR compliance as mentioned in the corresponding privacy policies. Building on previous research (Aberkane et al., 2022), this article aims to fill that gap by analyzing data processing entities' organizational factors to identify factors related to GDPR compliance promises in privacy policies.

Since privacy policies are typically articulated in textual form, this study aims to leverage their inherent linguistic nature effectively. Consequently, we employ a Natural Language Processing (NLP) based Machine Learning (ML) approach—ideally suited to dissect and understand these text-rich documents—to identify the aforementioned organizational factors related to GDPR compliance promises. Our research question (RQ) reads as follows: "What organizational factors are associated with the disclosure of GDPR compliance promises in the privacy policies of data processing entities?" Our contributions include 1) developing high-precision NLP-based classification models to evaluate GDPR compliance promises in privacy policies, 2) a comprehensive dataset of 8,614 EU organizations, including organizational information and GDPR compliance promises, and 3) an analysis of organizational factors that correlate with GDPR compliance promises.

This article starts in Section 2 by presenting relevant background information about the GDPR and the considered GDPR privacy policy core requirements. Section 3 follows this by discussing related works in the field. The adopted research methodology is then described in Section 4. Section 5 focuses on the developed classification models, followed by the data collection in Section 6. Section 7 presents the analysis of this work, leading to the discussion in Section 8. The article culminates in Section 9 with a conclusion of our research. Finally, we close by detailing the limitations of our work in Section 10, setting the stage for potential future research.

## 2. Background

This section provides an overview of the GDPR, the role of privacy policies in disclosing organizations' data processing practices, and the core GDPR privacy policy requirements considered in this article.

### *2.1. GDPR compliance & privacy policies*

The most relevant parts of the GDPR to this study are the regulations focusing on privacy disclosure and how these relate to privacy policies. In particular, we depart from Article 12, where the GDPR states that the controller—the person or entity that determines the purposes and means of the processing of personal data—should take appropriate measures to provide any information related to the processing of personal data, to the data subject in question in a "concise, transparent, intelligible and easily accessible form, using clear and plain language" (GDPR, 2016). In general, a privacy policy does precisely this. It discloses the data processing practices of the data processing entity (Karjoth & Schunter, 2002). For this reason, we consider privacy policies as the base for studying and analyzing the GDPR compliance promises of data processing entities.

### *2.2. GDPR privacy policy core requirements*

This research does not consider all possible GDPR compliance promises organizations make in their privacy policies. Instead, we scope our analysis to five core GDPR privacy policy requirements, considered generic and easily identifiable (Müller et al., 2019). This approach expands upon our previous study, which focused exclusively on the requirement of Purpose (Aberkane et al., 2022). The five core requirements included in this analysis are listed in Table 1.

Table 1. Five GDPR privacy policy core requirements (Müller et al., 2019)

| GDPR Requirement | Description | GDPR reference |
|---|---|---|
| Data Protection Officer | Notice of the contact details of the data protection officer, where applicable | Art. 13: §1b |
| Purpose | Notice of the purposes of the processing for which the personal data are intended | Art. 13: §1c |
| Acquired Data | Notice that personal data is, or is not, collected, and/or which data is collected | Art. 12,<br>Art. 14: §1d |
| Data Sharing | Notice of 3rd parties that can or cannot access a user's personal data | Art. 13: §1e, §1f |
| Rights | Notice of the user's right rectification and erasure | Art. 13: §2b |

A data processing entity must appoint a Data Protection Officer (DPO) if the core data processing activities involve processing sensitive personal data on a large scale (GDPR, 2016). Even if no legal obligation exists, designating a DPO or an equivalent role to lead the data processing activities in good channels and ensure compliance is recommended. Furthermore, among the information that has to be provided where personal data are collected from the data subject is Purpose. This comprises the purposes of the processing for which the personal data are intended and the legal basis for processing. Next, regarding Acquired Data, the data processing entity should give notice of the categories of personal data concerned. Also, it should be disclosed whether the data will be shared with third parties—in line with the requirement of Data Sharing. Lastly, the data processing entity must communicate the existence of the user's Rights, which are limited to the right to rectification and erasure in this study.

### *2.3. Context of the study*

In setting the context for our study, it is crucial to understand the potential causes, complexities, and implications associated with GDPR non-compliance, as these directly influence the GDPR compliance commitments stated in privacy policies. For example, small and medium-sized enterprises may struggle with resource constraints, complicating the implementation of GDPR's technical and organizational requirements (Freitas & Da Silva, 2018; Kapoor et al., 2018). Moreover, publicly listed companies face their own set of challenges concerning GDPR compliance. These companies operate in the public eye, with their operational activities and financial statements subject to public scrutiny (Ghonyan, 2017). As a result, GDPR non-compliance for such companies can have significant repercussions, including potential reputational damage (Ford, 2023).

GDPR compliance can also vary between countries and may be influenced by numerous factors, such as the country's regulatory environment and digital proficiency. Germany, for example, has demonstrated a longstanding commitment to regulating personal data processing, which is evident in its robust data privacy framework (Riccardi, 1983; Zell, 2014). This dedication is further evidenced by Germany's imposition of more fines than any other EU Member State during the initial year of GDPR enforcement (Barrett, 2020). The same is true for Sweden (Bygrave, 1998), potentially contributing to improved compliance practices today. However, a recent study argues that despite its historical precedent, Sweden still needs to meet GDPR standards (Herlin-Karnell, 2020), highlighting the ongoing complexities even for countries with established data protection histories. Additionally, the interpretation of the GDPR may vary between countries due to (subtle) language differences in its translations (Dexe et al., 2022), adding another layer of complexity to compliance efforts.

Furthermore, it is worth considering that the level of digital proficiency within a country—e.g., as per the European Commission's Digital Economy and Society Index (2022a)—can potentially impact its capacity to adhere to the GDPR effectively. Denmark, for instance, stands out as a digital front-runner in the EU and globally (European Commission, 2022b). Nevertheless, even with this seemingly favorable environment, interpretation of the GDPR remains a complex task (Motzfeldt & Næsborg-Andersen, 2022).

Moreover, the approach of countries can change over time. A noteworthy illustration of this is the Spanish Data Protection Agency, which appears to have adopted a stricter stance following the initial two years of GDPR enforcement (Levis & Fischer, 2021). This is suggested by the substantial number of fines imposed in Spain, which currently stands at 872 (CMS Law, 2024). In contrast, Eastern European countries have generally imposed fewer fines than their Western European counterparts (Daigle & Khan, 2020). For example, Slovenia, which only recently implemented the Data Protection Act aligning with the GDPR (Frantar & Gajšek, 2023; Vrabec, 2020), has yet to impose fines (CMS Law, 2024). However, it should be noted that the fines mentioned are only those that have been publicly disclosed, which means the actual numbers may differ.

## 3. Related work

GDPR-related research questions have been increasingly addressed with NLP techniques (Aberkane et al., 2021). This trend, however, does not include—to our knowledge—research conducted on identifying organizational factors using NLP and ML. In what follows, we briefly overview two adjacent research streams involving NLP and ML: privacy policy readability and the evaluation of privacy policy completeness, identified using the backward snowballing approach (Wohlin, 2014).

### *3.1. Privacy policy readability*

Several authors have addressed the readability of privacy policies in light of the GDPR. For example, Tesfay et al. (2018) outline the most relevant parts of the privacy policy using ML, employing PrivacyGuide, a privacy policy summarizing tool, classifying privacy policy content into eleven privacy aspects. Influenced by the GDPR, the authors aim to support Internet

users by simplifying the readability of privacy policies. Similarly, Zaeem et al. (2020) present PrivacyCheck v2. This ML-based tool automatically summarizes privacy policies by answering key questions, including questions that cover the essential concerns addressed by the GDPR. The tool aims to educate users on how their personal data is used on the Internet and how to select companies that attach more value to data protection.

### *3.2. Evaluating privacy policy completeness*

The majority of identified relevant literature focuses on evaluating privacy policy completeness in light of the GDPR. For instance, Liepina et al. (2019) designed a methodology for annotating post-GDPR privacy policies to identify and assess compliance with the GDPR using legal analysis, ML, and NLP—aiming to aid consumers with, among other things, understanding their rights and obligations as per the GDPR. More recently, Amaral et al. (2019) proposed AI-based automation for the completeness checking of privacy policies according to the GDPR—evaluated using 234 privacy policies from the fund industry—achieving precision and recall of 92.9% and 89.8%, respectively. Furthermore, using ML and rule-based analysis, Liu et al. (2021) propose an approach to analyze privacy policy contents and identify violations against Article 13 of the GDPR. Besides completeness, the authors also touch upon readability by implementing the approach in AutoCompliance, a web-based tool that reduces the user reading time by 55%. Along the same lines, El Hamdani et al. (2021) use rule-based approaches combined with ML to develop methods to automate compliance checking of privacy policies. In particular, the authors build a two-module system to verify the GDPR compliance of privacy policies, focusing mainly on the completeness of privacy policies. The first module extracts data practices, while the second module checks these extracted data practices on, among other things, the presence of mandatory information according to the GDPR. Lastly, Müller et al. (2019) introduce a data set of annotated privacy policies based on five GDPR privacy policy core requirements containing 18,397 natural sentences. The authors then proceed to design classifiers and evaluate the state of GDPR compliance "in the wild" by crawling privacy policies from actual companies. The results show that at least 76% of the privacy policies do not comply with at least one of the considered GDPR requirements.

### *3.3. Research gap*

In sum, contrary to previous literature, this study aims to not only investigate whether core requirements of the GDPR are disclosed in the privacy policy—thus slightly overlapping with the completeness research stream—but also aims to map the organizational factors that correlate with this disclosure. Expanding on previous research, this study identifies organizational factors (e.g., location, size, and sector) correlating with GDPR compliance, considering 8,614 data processing entities' privacy policies.

## 4. Methodology

This section describes the three-staged research methodology of this study.

### *4.1. Stage 1: training classification models*

The starting point is developing a supervised ML-based NLP pipeline to build five classifiers to assess whether privacy policies disclose the five GDPR privacy policy core requirements. This expands on our prior work, which developed a classifier for only one requirement. The data set for training the models was acquired from (Müller et al., 2019) containing 250 anonymized privacy policies comprising over 18,300 natural sentences, each labeled according to the GDPR privacy policy core requirements of Table 1. This stage resulted in five different classification models—trained using Python's Scikit-Learn library (Pedregosa et al., 2011)—each focusing on one of the requirements. Section 5 further elaborates on developing and evaluating the developed classification models.

### *4.2. Stage 2: data gathering & classification*

The second stage focused on collecting a fruitful data set for analysis due to the absence of organizational information related to the privacy policies in the data set utilized in the previous stage. Moreover, the data had been anonymized, precluding the possibility of identifying the company that authored each privacy policy and collecting its relevant organizational data. Thus, utilizing Bureau van Dijk's Orbis database, we gathered data from 168,824 Europe-based companies (Bureau van Dijk, 2021). Following this, we scraped—if possible—the publicly accessible privacy policies of each company, resulting in 10,090 policies (see Section 6 for details). Subsequently, these policies were analyzed using five classification models to determine the disclosure of the five GDPR core requirements. The classification results were then combined with the organizational data of the related company into one data set. Finally, this combined data set was filtered from irrelevant data in anticipation of the final analysis of stage three, resulting in a data set containing the organizational data and classification results of 8,614 companies.

### *4.3. Stage 3: analysis*

The data set constructed in Stage 2 was analyzed to assess to what extent the organizational factors are associated with the disclosure of the five GDPR privacy policy core requirements. This analysis used a separate logistic regression model, utilizing Python's Statsmodels library (Seabold, 2010). The results are presented in Section 7.

## 5. Classification

The data set for training our models, collected and annotated by Müller et al. (2019), includes 18,397 sentences from 250 privacy policies. The authors amassed the data by scraping the privacy policies followed by manual annotation according to the following five GDPR privacy policy core requirements: DPO, Purpose, Acquired Data, Data Sharing, and Rights. Table 2 presents statistics related to each GDPR class (interchangeably used with GDPR privacy policy core requirements in the remaining sections), showing the sentence counts per class and the average sentence counts per policy.

A sentence is considered to comply with the DPO privacy policy core requirement if "the Data Protection Officer or an equivalent authority is named, or contact details of a similar authority are provided" (Müller et al., 2019). The Purpose requirement is met if the processing purposes are disclosed. The Acquired Data requirement is fulfilled when the collected data is specified. The Data Sharing requirement is satisfied by disclosing information about personal data sharing. Lastly, the authors of the data set limited the scope of the Rights requirement to two GDPR instances: the right to be forgotten and the right to rectification.

Table 2. Data set statistics

| GDPR Class | Number of sentences in corpus | Average number of sentences per privacy policy |
|---|---|---|
| Data Protection Officer | 414 | 2 |
| Purpose | 980 | 4 |
| Acquired Data | 565 | 3 |
| Data Sharing | 830 | 4 |
| Rights | 251 | 2 |

### *5.1. Training the model*

This section covers the six steps involved in training and evaluating the classification model, as depicted in Fig. 1.

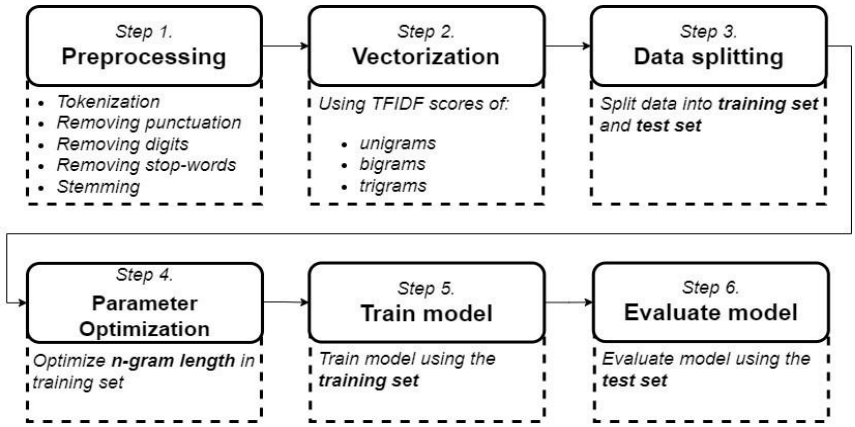



Fig. 1. Process of training and evaluating the classification models

**Step 1.** First, the corpus was led through a traditional NLP pipeline. This preprocessing process consisted of tokenization, punctuation removal, digit removal, and stemming.

**Step 2.** Subsequently, vectorization of the sentences took place based on the Term Frequency-Inverse Document Frequency (TFIDF) (Robertson, 2004). In particular, as features for our classification model, we utilized the TFIDF scores of different modes of n-grams (i.e., sequences of tokens of length $n$): unigrams, bigrams, and trigrams.

**Step 3.** The data was then divided into training and test sets, increasing the test set size from 0.1 (in our preliminary study) to 0.2. Furthermore, as shown in Table 2, the data set is imbalanced. In fact, over 83% of the sentences are not labeled with one of the five GDPR privacy policy core requirements. To address this imbalance, we conducted a stratified split of the data—for each of the five GDPR privacy policy core requirements (i.e., GDPR classes)—into a training and test set, followed by an oversampling of the training set. The stratified split maintained the ratio between positive and negative sentences, while the oversampling increased the representation of the minority class (i.e., the class of interest).

**Step 4.** Parameter optimization was performed based on the Area Under the Receiver Operating Characteristic Curve (ROC AUC) using 5-fold cross-validation. The aim was to identify the classification model and parameter settings (mainly focusing on the n-gram length) that yield the most promising results. This resulted in settling on logistic regression, an established and appropriate technique for the addressed problem, i.e., a supervised binary classification problem.

**Step 5.** This step comprised training the classification models using the optimized n-gram configuration. However, the addressed problem can also be interpreted as a multi-label problem, as the labels are not mutually exclusive. Because of the latter, we opted to address each class individually for simplicity regarding the issue of class imbalance, resulting in five different binary classification problems. Hence, we trained five logistic regression models utilizing the scikit-learn library—configured with the optimized parameters—on the prepared data set.

**Step 6.** The evaluation of the classification models took place using the test set. The performance of the five classification models is presented in Table 3. We utilized one of the most often used performance metrics for binary classification: the ROC AUC score (Sokolova & Lapalme, 2009). The ROC AUC score represents a measure of the ability of a classifier to distinguish between classes. It corresponds with the probability that the model will score a randomly chosen positive instance higher than a random negative one. A score of 1 corresponds with a perfect model, and 0.5 corresponds with a random model.

Table 3. Initial sentence classification performance

| GDPR Class | ROC AUC score |
|---|---|
| Data Protection Officer | 0.965 |
| Purpose | 0.848 |
| Acquired Data | 0.856 |
| Data Sharing | 0.921 |
| Rights | 0.950 |

### *5.2. Calibration of precision*

Given that the models predict at a sentence level, we need to address the following question: *When does a privacy policy meet the GDPR privacy policy core requirement at issue?* Is the mere presence of one positively classified sentence enough to consider the whole privacy policy as disclosing the requirement in question? We decided to raise the certainty of our predictions by addressing the following question: *What number of positive sentences are needed to classify, with a desired level of confidence, a new privacy policy as disclosing the requirement at issue?* We used the inverse cumulative distribution function of the binomial distribution to set a threshold of the minimum positive sentences required to meet a given confidence level. This step is necessary since our classification models work at the sentence level rather than the document level due to the data set used to train the models. This data set consists of labeled sentences extracted from privacy policies without insight into the structure of the original documents (i.e., privacy policies).

We use a contingency table consisting of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values based on a data set comprising, in total, $\mathrm{E} = \mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}$ elements. Note that these values were derived using an optimized cut-off point based on the F1-Score to rectify the bias that permeated the classification model due to oversampling. In this context, $\hat{P}$ describes the elements predicted as positive, meaning $\hat{P} = TP + FP$. The probability that a positively predicted element is a true positive can then be described as $P(y = 1|\hat{y} = 1) = \frac{TP}{\hat{P}}$. Using a binomial distribution, the probability that exactly $k$ elements from $\hat{P}$ are true positives can be calculated as follows:

$$P(|TP| = k) = \binom{\hat{P}}{k} P(y = 1|\hat{y} = 1)^k \big(1 - P(y = 1|\hat{y} = 1)\big)^{\hat{P}-k} = \binom{\hat{P}}{k} P_{TP}^k P_{FP}^{\hat{P}-k} \quad (1)$$

It follows that the cumulative distribution function is equal to the following:

$$P(|TP| \leq k) = \sum_{i=0}^{k} \binom{\hat{P}}{i} P_{TP}^k P_{FP}^{\hat{P}-i} \quad (2)$$

We focus on the probability that the number of TP exceeds a given value. Therefore, we consider the inverse cumulative distribution:

$$P(|TP| > k) = 1 - P(|TP| \leq k) = 1 - \sum_{i=0}^{k} \binom{\hat{P}}{i} P_{TP}^k P_{FP}^{\hat{P}-i} \quad (3)$$

We now find the highest value of $k$ (i.e., $k'$) that keeps the inverse cumulative distribution above a given probability $Z$:

$$k' = argmax_{k \in [0,|TP|]} P(|TP| > k) \geq Z \quad (4)$$

Given a new set of elements (e.g., sentences in a privacy policy) with $\mathrm{E_n}$ elements of which $\widehat{\mathrm{P_n}}$ are predicted as positive by the model. We consider the document to be positive if the threshold, $\mathrm{Q} = \frac{\mathrm{k'}}{\mathrm{E}}$ is met, i.e., $\frac{\widehat{\mathrm{P}}}{\mathrm{E_n}} \geq \mathrm{Q}$. Table 4 presents the calibrated thresholds for all GDPR classes, combined with the respective precision.

Table 4. Document classification performance

| GDPR Class | Threshold $Q$ | Precision per document |
|---|---|---|
| Data Protection Officer | 0.016 | 0.908 |
| Purpose | 0.028 | 0.908 |
| Acquired Data | 0.016 | 0.928 |
| Data Sharing | 0.030 | 0.912 |
| Rights | 0.009 | 0.941 |

### *5.3. Example*

Consider the DPO class, for example, where the threshold $\mathrm{Q}$ has been calibrated using a test set size of 3,680 ($\mathrm{E_{test}}$), while the desired probability (i.e., $\mathrm{Z}$) was set at 90%. The calibration result shows that for a given new set of elements (e.g., sentences of a privacy policy), the following threshold $\frac{\widehat{\mathrm{P}}}{\mathrm{E_n}} \geq 0.016$ must be met to achieve a precision of at least 0.908. Given a privacy policy comprising 100 sentences (i.e., $\mathrm{E_n} = 100$), we arrive at the following equation:

$$\frac{\widehat{\mathrm{P}}}{100} \geq 0.016 \quad (5)$$

Then, to classify the privacy policy—with a precision of at least 0.908—we require $\widehat{\mathrm{P}} \geq 1.6$. Since we speak in terms of sentences and not sentence fragments, we round the value of required positively predicted sentences ($\widehat{\mathrm{P}}$) up to 2. In sum, to classify a privacy policy containing 100 sentences as meeting the DPO requirement, at least two of its sentences must be predicted into the GDPR class of DPO.

### *5.4. Tools*

The classification models were developed and evaluated in *Python 3.8.5*. The preprocessing and vectorization (using TFIDF (Robertson, 2004)) steps were conducted using Python's *Natural Language Toolkit (NLTK)* (Bird et al., 2009) library. The tokens were stemmed using *PorterStemmer*. The splitting of the data, parameter optimization, training, and model evaluation was managed by Python's *Scikit-Learn library* (Pedregosa et al., 2011). The oversampling of the imbalanced (training) data was performed using *imbalanced-learn's RandomOverSampler* technique (Lemaître et al., 2017).

## 6. Data collection

The data collection process consists of two stages: collecting organizational data from Orbis and scraping the privacy policies of the corresponding organizations. We will delve into the details of these stages in what follows.

### *6.1. Collecting organizational data*

We used the Orbis database (Bureau van Dijk, 2021) to collect organizational data from organizations the GDPR applies to. From this database, we collected a random sample of organizational data of 168,824 companies located in the EU, including the following organizational details: the **company name**, **quoted** (describing whether the company was publicly listed), the **country ISO code** indicating the location of the company, **NACE code** depicting the business activity of the company (i.e., sector), the **last available year** of the data, the **operating revenue** based on the last available year, the **number of employees**, and the **size classification** of the organization.

However, not all organizational factors were included in our final set of factors considered for analysis (Table 5). The **company name** was used to crawl the web and scrape the related privacy policy, but it was omitted afterward. Similarly, **the last available year** was only used to sift the data and keep only the data applicable to the GDPR, i.e., data published or updated after the enactment of the GDPR in 2018. Also, we shifted from Orbis' categorization of size, i.e., *small companies*, *medium-sized companies*, *large companies*, and *very large companies*). For the sake of simplicity, *small companies*, *medium-sized companies* were united under the umbrella of "small and medium-sized enterprises", whereas *large companies*, and *very large companies*) were combined into "large enterprises." The latter, i.e., "large enterprises," is defined, according to Orbis, by at least one of the following three criteria: (1) an operating revenue equal to, or more than, 10 million euros, (2) owning total assets equal to, or more than, 20 million euro, and (3) having 150 employees or more. If a company does not meet these conditions, it is categorized under "small and medium-sized enterprises."

Table 5. The final set of organizational factors considered for analysis

| Name | Description |
|---|---|
| Country ISO code | *ISO 3166-1 alpha-2: two-letter country code.* |
| NACE Rev. 2 code (level 1) | *Classification of 21 business activities, e.g., "agriculture, forestry, and fishing."* |
| Number of employees | *Number of employees as reported in the last available year.* |
| Operating revenue | *Operating revenue as reported in the last available year.* |
| Quoted | *Boolean value indicating whether the company is listed or not.* |
| Size classification | *Boolean value describing the company's size (i.e., small and medium-sized or large enterprise).* |

### *6.2. Scraping privacy policies*

After collecting a sample of organizational data of 168,824 companies located in the EU, we aimed to gather the related privacy policies of the organizations. To do so, we devised a web-scraping algorithm in Python—using the *urllib*, *google-search* (Vilas, 2020), *Newspaper* (Ou-Yang, 2013), and NLTK libraries—identifying and collecting the relevant privacy policies resulting from querying *Google Search*. The privacy policies were collected over 20 days in September 2021. The web scraper followed the following steps:

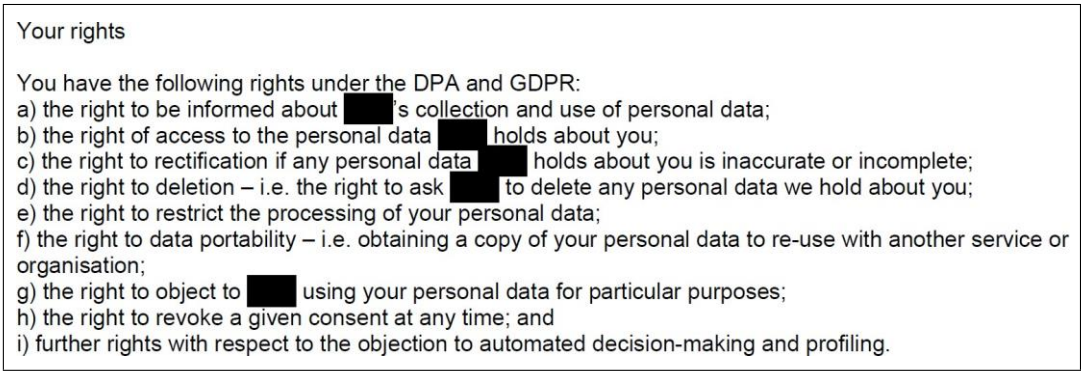
**Step 1.** Query Google Search with the company name and attempt to identify in the first three results whether (parts of) the company name is present in the Uniform Resource Locator (URL) and, more specifically, in the network location part (netloc) of the URL, according to the general structure of a URL: scheme://netloc/path;parameters?query#fragment.

**Step 2.** If the netloc included (parts of) the company name, the URL was saved for the next step. If not, the URL was skipped, and the following URL from the Google Search results was considered, with a maximum of three attempts. Finally, the company at issue was omitted from the data set if there was no positive result.

**Step 3.** Using the relevant netloc, we devised the following Google query to search the specific website of the affiliated organization for a privacy policy: "site:" + <netloc> + " privacy policy".

**Step 4.** Next, three attempts were made to scrape the relevant privacy policy. The URLs that resulted from the query were examined. If the term "privacy" or "policy" was present in the URL, the text on the corresponding page was scraped *if* the text was written in English *and* longer than ten sentences. The latter requirement was set in place to avoid irrelevant pages. If these requirements were not met, the next URL in line was examined. This process was repeated up to three times.

We were able to scrape 10,090 privacy policies with this process. Fig. 2 shows the disclosure of GDPR compliance promises in one of the scraped privacy policies. These privacy policies were then classified into the five GDPR privacy policy core requirements using the classification models of Section 5. Next, the classification results were tied with the related organizational data from the initial Orbis data set. Finally, this combined data set was sifted from irrelevant data, i.e., data published or updated before the enactment of the GDPR in 2018, resulting in a data set containing organizational data of 8,614 companies, including the corresponding classification results.


Your rights

You have the following rights under the DPA and GDPR:
a) the right to be informed about [redacted]'s collection and use of personal data;
b) the right of access to the personal data [redacted] holds about you;
c) the right to rectification if any personal data [redacted] holds about you is inaccurate or incomplete;
d) the right to deletion – i.e. the right to ask [redacted] to delete any personal data we hold about you;
e) the right to restrict the processing of your personal data;
f) the right to data portability – i.e. obtaining a copy of your personal data to re-use with another service or organisation;
g) the right to object to [redacted] using your personal data for particular purposes;
h) the right to revoke a given consent at any time; and
i) further rights with respect to the objection to automated decision-making and profiling.

Fig. 2. Screenshot of GDPR compliance promises disclosed in one of the scraped privacy policies

## 7. Analysis

In this section, we detail how the collected data set—containing the organizational factors and GDPR compliance promises of 8,614 organizations—is prepared and then analyzed to identify the organizational factors associated with the GDPR privacy policy core requirements disclosed in each organization's privacy policy. The organizational factors considered for analysis were already outlined in Table 5. The full output of the analysis is made available at https://aberkane.github.io/GDPR-privacy-policies/.

### *7.1. Training the model*

**Step 1.** The data types of the predictors (i.e., organizational factors) and the target values (i.e., the GDPR privacy policy core requirements) were transformed into a categorical and numerical representation.

**Step 2.** The categorical data was encoded, whereas the numerical data was scaled.

**Step 3.** The data was split into a training set and a test set.

**Step 4.** Parameter optimization was performed by tuning the logistic regression parameters to maximize accuracy using the training set, and then evaluating performance with the test set. Accuracy describes the fraction of correct predictions over the total number of predictions: $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. The outcome of the parameter optimization is presented in Table 6, including regularization, alpha (i.e., the weight multiplying the regularization penalty term), and the resulting accuracy.

**Step 5.** The model was retrained on the entire data using the optimized parameters, applying a significance level of 0.05. The results are presented in the remainder of this section.

Table 6. Optimized parameters and corresponding accuracy for Statsmodels' logistic regression

| Name | Regularization | Alpha | Accuracy |
|---|---|---|---|
| Data Protection Officer | L1 | 8.101 | 0.636 |
| Purpose | L1 | 0.001 | 0.867 |
| Acquired Data | L1 | 5.501 | 0.829 |
| Data Sharing | L1 | 9.701 | 0.752 |
| Rights | L1 | 0.001 | 0.771 |

### *7.2. Results*

Table 7 summarizes the significant predictors and their coefficients associated with the GDPR privacy policy core requirement of **DPO**, which includes providing the contact information for the DPO or equivalent. The results show that being quoted positively correlates with the GDPR requirement of DPO, i.e., a publicly listed company is less prone to communicate the contact details of the DPO—where applicable—than an organization that is not listed. Furthermore, the location of the company headquarters also plays a role, with companies located in Italy having a lower probability of disclosing information regarding the DPO. In contrast, companies in Belgium, Germany, Denmark, France, Ireland, and Sweden are likelier to disclose information about the DPO.

Table 7. Significant predictors and corresponding coefficients for target value DPO

| Predictor | P-value | Coefficient |
|---|---|---|
| Quoted | $1.396142 \times 10^{-4}$ | 0.581482 |
| BE (Country ISO code) | $7.610162 \times 10^{-3}$ | 0.337784 |
| DE (Country ISO code) | $3.292703 \times 10^{-3}$ | 0.242856 |
| DK (Country ISO code) | $1.363752 \times 10^{-3}$ | 0.611399 |
| FR (Country ISO code) | $9.424592 \times 10^{-4}$ | 0.285202 |
| IE (Country ISO code) | $3.915174 \times 10^{-2}$ | 0.290846 |
| IT (Country ISO code) | $8.796016 \times 10^{-12}$ | −0.522975 |
| SE (Country ISO code) | $1.176022 \times 10^{-3}$ | 0.352673 |

Table 8 presents the significant predictors and their respective coefficients in relation to the GDPR privacy policy core requirement of **Purpose**, i.e., the disclosure of the purposes of the processing. Similar to the previous GDPR requirement of DPO, the results show that being quoted positively correlates with Purpose. Furthermore, the results reveal a positive

correlation between Purpose and Greece as headquarters locations of data processing entities. On the contrary, Slovenia negatively correlates with the Purpose requirement. Similarly, centering on the classification of business activities, "agriculture, forestry and fishing" and "public administration and defense; compulsory social security" negatively correlate with complying with the Purpose-criterion. Finally, relating to the size classification, the "small and medium-sized enterprises" predictor also shows a negative correlation with disclosing the purposes of the processing.

Table 8. Significant predictors and corresponding coefficients for target value Purpose

| Predictor | P-value | Coefficient |
|---|---|---|
| Quoted | 0.031873 | 0.576 070 |
| GR (Country ISO code) | 0.011031 | 1.031 066 |
| SI (Country ISO code) | 0.016175 | –0.904 906 |
| Agriculture, forestry and fishing (NACE Rev. 2 code) | 0.004734 | –1.046 389 |
| Public administration and defense; compulsory social security (NACE Rev. 2 code) | 0.040562 | –1.388 573 |
| Small and medium-sized enterprises (Size classification) | 0.001417 | –0.219 327 |

Next, Table 9 shows the corresponding statistically significant predictors and coefficients related to the GDPR privacy policy core requirement of **Acquired Data**: communicating whether (and which) personal data is or is not collected. The results show that only four predictors are significant predictors, all of which negatively correlate with the Acquired Data requirement. Notably, the negative correlation of the *quoted* predictor shows that publicly listed companies are less likely to comply with the Acquired Data requirement, diverting from the trend between *quoted* and the previous two GDPR requirements: DPO and Purpose. Similarly, companies headquartered in Spain, Italy, and Poland negatively correlate with the Acquired Data requirement.

Table 9. Significant predictors and corresponding coefficients for target value Acquired Data

| Predictor | P-value | Coefficient |
|---|---|---|
| Quoted | $5.280495 \times 10^{-16}$ | –1.185473 |
| ES (Country ISO code) | $3.542255 \times 10^{-20}$ | –0.993589 |
| IT (Country ISO code) | $1.110738 \times 10^{-9}$ | –0.606650 |
| PL (Country ISO code) | $2.841021 \times 10^{-4}$ | –0.532567 |

In the context of the GDPR privacy policy core requirement of **Data Sharing**, Table 10 presents significant predictors and their associated coefficients. The findings suggest that being quoted negatively correlates with compliance with the Data Sharing requirement, which is consistent with the previous GDPR requirement of Purpose. The results further indicate that four significant predictors in the *country ISO code* category positively correlate with the Data Sharing target value, specifically Denmark, Ireland, the Netherlands, and Sweden. Conversely, Spain shows a negative correlation with the Data Sharing target value. Finally, being categorized as a small or medium-sized enterprise negatively correlates with the Data Sharing requirement.

Table 10. Significant predictors and corresponding coefficients for target value Data Sharing

| Predictor | P-value | Coefficient |
|---|---|---|
| Quoted | $2.566122 \times 10^{-14}$ | −1.065388 |
| DK (Country ISO code) | $2.015918 \times 10^{-2}$ | 0.510705 |
| ES (Country ISO code) | $3.572360 \times 10^{-4}$ | −0.336990 |
| IE (Country ISO code) | $4.081384 \times 10^{-3}$ | 0.496183 |
| NL (Country ISO code) | $7.944794 \times 10^{-4}$ | 0.527240 |
| SE (Country ISO code) | $4.717011 \times 10^{-2}$ | 0.236954 |
| Small and medium-sized enterprises (Size classification) | $-4.825884 \times 10^{-4}$ | −0.186945 |

Finally, table 11 illustrates the findings concerning the GDPR's **Rights** privacy policy core requirement, which mandates the specification of the user's right to rectification and erasure. Consistent with all GDPR privacy policy core requirements analyzed, being quoted emerged as a significant predictor of the target value. Specifically, the *quoted* predictor positively correlated with the Rights requirement. Additionally, companies headquartered in Germany and Denmark positively correlated with the Rights requirement, whereas Croatia and Slovenia exhibited a negative correlation.

Table 11. Significant predictors and corresponding coefficients for target value Rights

| Predictor | P-value | Coefficient |
|---|---|---|
| Quoted | 0.000186 | 0.709050 |
| DE (Country ISO code) | 0.001392 | 0.592737 |
| DK (Country ISO code) | 0.014106 | 0.692819 |
| HR (Country ISO code) | 0.020157 | −0.625037 |
| SI (Country ISO code) | 0.017338 | −0.792389 |

## 8. Discussion

This section discusses the notable results of the analysis, as presented in the previous section. In general, we focus on the predictors that show a significant correlation with two or more target values. It is worth noting that, except for the *quoted* predictor, the results did not show ambiguity, as no predictor had both a positive and negative correlation with the GDPR privacy policy core requirements.

First, the results show that being a publicly listed company—which corresponds to the **quoted** predictor—correlates with all five considered GDPR privacy policy core requirements. DPO, Purpose, and Rights correlate positively, while Acquired Data and Data Sharing correlate negatively. A publicly listed company usually discloses information where the operational activities and financial statements are public (Ghonyan, 2017), which might increase the need to comply with the GDPR, as non-compliance might lead to reputational damage. Even a small fine can significantly impact the market value (Ford, 2023). Moreover, listed companies typically have more resources since, for example, the organization's net value increases, and the debt-to-equity ratio improves by going public (Ghonyan, 2017). These resources can aid in meeting the requirements of the GDPR, as they allow for carrying out the necessary measures to comply with the GDPR requirements. Therefore, the positive correlations are in agreement with prior studies. However, despite all of the above, results show a negative correlation between a publicly listed company and the disclosure of the nature of personal data collected

(Acquired Data) *and* disclosing whether third parties can or cannot access personal data (Data Sharing), revealing room for improvement.

Centering on **small and medium-sized enterprises**, the results reveal a negative correlation with the GDPR privacy policy core requirements of Purpose and Data Sharing. This result is in line with previous research where the issue of resource poverty regarding small and medium-sized enterprises is pointed out: complying with the requirements of the GDPR may be problematic for small and medium-sized enterprises as they might struggle with taking the required technical and organizational steps towards GDPR compliance (Freitas & Da Silva, 2018; Kapoor et al., 2018) The high cost of non-compliance, including hefty fines, can also pose a significant burden on small and medium-sized enterprises, further exacerbating the difficulties they face in meeting the requirements of the GDPR. The predictor was not found to be significant in the remaining classes.

Regarding organizations' headquarters, the results indicate that **Denmark** positively correlates with the GDPR privacy policy core requirements of DPO, Data Sharing, and Rights, making it the best-performing host country. This favorable performance of Denmark concerning the disclosure of GDPR compliance requirements may be attributed to its achievements in the digital sphere. According to the Digital Economy and Society Index, which presents countries' performance in digitization (European Commission, 2022a), Denmark is "a digital front-runner both in the EU and globally and continues to progress relatively well." Furthermore, Denmark is ranked first in the EU in the connectivity dimension: 95% of households are connected to very-high-capacity networks (European Commission, 2022b). This progress and attitude are reflected in the GDPR measures taken after the looming of the GDPR. Denmark has set up the (already existing) Data Protection Agency to promote awareness of the GDPR and ensure it is followed. The Danish Data Protection Agency also provides guidance and support to companies and individuals so that they can comply with the GDPR.

Its Scandinavian counterpart, **Sweden**, positively correlates with the target values of DPO and Data Sharing. Moreover, like Denmark, Sweden is considered one of the digital front-runners in the EU (European Commission, 2022a). These positive correlations may be attributed to the early adoption of data protection statutes (Bygrave, 1998). Nonetheless, despite this early development, Herlin-Karnell (2020) argues that there is, in fact, a lack of compliance in Sweden due to a "disproportionate reading of the GDPR", mentioning that private actors may purchase a publishing license that exempts them from the GDPR, allowing them to share information about individuals and even earn a profit from this under the banner of freedom of expression. However, to validate the claim indicating that the Swedish reading of the GDPR gives room for actions that may be classified as non-compliance in other EU countries, a thorough investigation is needed into the data processing activities of organizations in Sweden, comparing these activities to the GDPR requirements.

Along the same lines, the results indicate that **Ireland**—which acts as a host for major U.S. digital services providers—positively correlates with the GDPR privacy policy core requirements of DPO and Data Sharing. The country's low corporate tax regime (Gunnigle & McGuire, 2001) and its status as a center for multinational corporations make it an attractive location for companies to establish their European operations. These companies must comply with the GDPR, which might explain the country's high level of compliance. Nevertheless, Daigle and Khan collect some skepticism towards the popularity of Ireland as a headquarters location for the EU operations of major digital services providers, mentioning the concern shared by some EU Member States' Data Protection Authorities about Ireland possibly permitting significant violations of the General Data Protection Regulation (Daigle & Khan, 2020). In 2023, however, the Irish Data Protection Commission (DPC) issued a 1.2 billion euro fine to Meta Platforms Ireland Limited (Meta IE) (EDPB, 2023), challenging the allegations (Francis, 2022).

Another country where the results were optimistic is **Germany**, which is positively associated with the GDPR privacy policy core requirements of DPO and Rights. Germany's commitment to GDPR implementation is demonstrated by the fact that it imposed more fines than any other EU Member State in the first year of enforcement, highlighting its dedication to data privacy regulation (Barrett, 2020). In contrast, almost half of the EU Member States did not issue fines in the first year, many due to a lack of sufficient resources. This decisiveness was to be expected since, even before the advent of the

GDPR, Germany upheld one of the most robust data privacy protection frameworks in the world (Zell, 2014), grounded on, among other acts, the German Federal Data Protection Act of 1977, which protects all forms of personal data on an identified or identifiable natural person, setting a "commendable" precedent in regulating the processing of personal data (Riccardi, 1983). Therefore, it can be argued that Germany's promising results might be linked to its strong commitment to data privacy protection and its pre-existing robust data protection framework.

On the other side of the fence, the results indicate that data processing organizations headquartered in **Spain** are less likely to disclose information regarding Acquired Data and Data Sharing. It might be due to these reasons that the Spanish Data Protection Agency, Agencia Española de Protección de Datos, seems to have taken a more demanding approach after the first two years of GDPR enforcement (Levis & Fischer, 2021). Furthermore, according to the CMS Law GDPR Enforcement Tracker, Spain collected, by far, the most fines since the enforcement of the GDPR (CMS Law, 2024). At the time of writing, the number of fines accumulated to 872, which is more than twice the fines imposed by the second country on the list, Italy, with 377 fines. **Italy**, as a predictor, negatively correlates with the target values DPO and Acquired Data. The Garante per la Protezione dei Dati Personali, i.e., the Italian Data Protection Authority, was slower than most other EU countries in implementing fines against companies following the introduction of the GDPR (Daigle & Khan, 2020). However, beginning in early 2020, it began issuing more severe fines. As a result, and as previously noted, the authority has accumulated a total of 377 fines as of the time of writing. The number of fines imposed by both Spain and Italy seems to agree with the narrative of the results. However, it is important to contextualize these figures. The issuance of a large number of fines does not necessarily indicate an excessive level of non-compliance with the GDPR. Instead, it could indicate the diligent enforcement of the regulation by the authorities. Additionally, it is noteworthy that Italy and Spain have among the largest staffs to support their respective Data Protection Authorities, which may contribute to the high number of fines (Barrett, 2020). This is because they are likely more capable of processing complaints efficiently and enforcing the GDPR than countries with limited resources. On the other hand, authorities with fewer resources might conserve their resources by prioritizing organizations handling sensitive personal data (Presthus & Sønslien, 2021).

Next, **Croatia** demonstrated a negative correlation with one of the requirements: organizations headquartered in Croatia are less likely to disclose information regarding rights in their privacy policies. Its Eastern European counterpart, **Slovenia**, negatively correlates with the requirements of Purpose and Rights. Daigle and Khan (2020) provide insight into the fining of Eastern European countries, thus including Croatia and Slovenia. The authors report that, from May 2018 to March 2020, 11 penalties out of a total of over 100 across the EU exceeded one million euros each, with only one of these 11 fines imposed by an Eastern European country. Furthermore, the authors note that of the 107 fines exceeding 10,000 euros each issued between May 2018 and March 2020, only 24 percent originated in Eastern European countries, possibly indicating a more lenient approach. Focusing on Croatia's fining pattern, we find that out of a total of 29 fines, the vast majority (25) were imposed after 2021 (CMS Law, 2024), indicating a trend towards more stringent GDPR enforcement. In contrast, Slovenia has been slower in complying with GDPR standards. Despite the regulation being instated in the EU in 2018, the Slovenian Parliament only adopted the Data Protection Act (ZVOP-2), a national law implementing the GDPR, on 15 December 2022 (Frantar & Gajšek, 2023; Vrabec, 2020). Currently, no fines have been imposed by the Slovenian Data Protection Authority (CMS Law, 2024). It is worth noting that the fines referenced are based on publicly disclosed data, which may not fully represent the actual figures.

Regarding the industry classification of **NACE**, the results indicate that, among the 21 industries, only two industries showed a significant correlation: "agriculture, forestry and fishing" and "public administration and defense; compulsory social security" negatively correlated with the target value of Purpose. These results imply that the industry classification of data processing entities may not be a reliable predictor of their level of GDPR compliance, as revealed in their privacy policies. The same applies to the predictors of **number of employees** and **operating revenue**, as both predictors demonstrated insignificance.

In conclusion, the findings regarding the predictor of **quoted** lack consistency, as negative and positive correlations were detected. However, the results of this study seem to support the notion that small and medium-sized enterprises face difficulties complying with the GDPR due to limited resources. Additionally, initial observations may suggest a potential relationship between a country's level of maturity in data protection legislation and the disclosure of GDPR privacy policy core requirements by data processing companies headquartered in that country. Finally, previous research suggests that Eastern European Member States may have a distinct approach and attitude toward GDPR compliance compared to their Western counterparts (Daigle & Khan, 2020; Vrabec, 2020). However, validation of these possible causal links requires further research. An overview of all significant correlations—where "+" signifies a positive correlation and "-" signifies a negative correlation—is presented in Table 12.

Table 12. Overview of all significant correlations between the organizational factors and the GDPR privacy policy core requirements

| Organizational Factor | | GDPR Privacy Policy Core Requirements | | | | |
|---|---|---|---|---|---|---|
| | | DPO | Purpose | Acquired Data | Data Sharing | Rights |
| Country ISO Code | *BE (Belgium)* | + | | | | |
| | *HR (Croatia)* | | | | | − |
| | *DK (Denmark)* | + | | | + | + |
| | *FR (France)* | + | | | | |
| | *DE (Germany)* | + | | | | + |
| | *GR (Greece)* | | + | | | |
| | *IE (Ireland)* | + | | | + | |
| | *IT (Italy)* | − | | − | | |
| | *NL (The Netherlands)* | | | | + | |
| | *PL (Poland)* | | | − | | |
| | *SI (Slovenia)* | | − | | | − |
| | *SP (Spain)* | | | − | − | |
| | *SE (Sweden)* | + | | | + | |
| NACE Rev. 2 code | *Agriculture, forestry, and fishing* | | − | | | |
| | *Public administration and defense; compulsory social security* | | − | | | |
| Number of employees | | | | | | |
| Operating revenue | | | | | | |
| Quoted | | + | + | − | − | + |
| Size | *SME* | | − | | − | |
| | *LE* | | | | | |

## 9. Conclusion

In this study, we utilized ML and NLP to address the research question: *"What organizational factors are associated with the disclosure of GDPR compliance promises in the privacy policies of data processing entities?"* The study focused on five GDPR privacy policy core requirements disclosed in privacy policies and examined six organizational factors and their subclasses for their effect on the disclosure of these requirements.

The study made several contributions. Firstly, we developed five NLP-based classification models with precision scores of at least 0.908 to evaluate GDPR compliance promises revealed in privacy policies. This approach offers researchers an innovative method for efficiently and accurately analyzing large volumes of privacy policies, with its calibration step ensuring high precision. Secondly, a data set of 8,614 organizations in the European Union was compiled, comprising organizational information and the GDPR compliance commitments extracted from the privacy policy of each organization. This dataset can serve as a valuable resource for researchers conducting further studies on GDPR compliance and for professionals benchmarking their compliance efforts against a broad range of peers. Lastly, we analyzed the organizational factors that correlated with the disclosure of GDPR compliance promises in privacy policies. These insights can guide future research by helping identify issues and patterns in GDPR compliance across various organizations. Moreover, they can aid professionals in developing targeted approaches to enhance GDPR compliance. For example, our analysis reveals that SMEs are less likely to disclose certain GDPR requirements in their privacy policies, suggesting that these organizations may need additional support. Using this insight, professionals such as policymakers or governmental actors can devise specific training programs or resources tailored to SMEs to enhance their compliance efforts effectively.

Our findings demonstrated that being a small or medium-sized enterprise negatively correlated with disclosing two GDPR privacy policy core requirements. The findings regarding the location of data processing entities revealed 13 significant predictors. Eight countries—Denmark, Germany, Ireland, Belgium, France, Greece, the Netherlands, and Sweden—positively correlated with one or more GDPR privacy policy core requirements without negatively correlating with the remaining requirements. On the other hand, Poland, Croatia, Spain, Italy, and Slovenia negatively correlated with the GDPR privacy policy core requirements. When analyzing the target values, it was observed that the predictors generally tended to correlate positively with DPO and negatively with Purpose and Acquired Data. For Data Sharing and Rights, the number of positive and negative correlations was more balanced.

The study results indicate a potential relationship between a country's level of maturity regarding data protection legislation and the disclosure of GDPR privacy policy core requirements by data processing companies headquartered in that country. Similarly, the results suggest a possible relationship between a country's level of digitization and the disclosure of these core requirements.

This study adds to the growing body of research on GDPR, providing new insights into the challenges of compliance with this regulation. The results were contextualized by aligning them with available literature and statistics. These contributions serve as a foundation for further academic research on the issue of non-compliance. However, the potential of these findings reaches beyond academia, offering a robust basis for professionals.

## 10. Limitations and future work

This research focuses on five GDPR privacy policy core requirements, although the GDPR encompasses more comprehensive regulations. For instance, in this study, the "Right" requirement is limited to the right to rectification and erasure, whereas the GDPR also includes other rights, such as the right to data portability mentioned in Article 20 (GDPR, 2016). We scoped the research to key GDPR requirements that are easily inferable from privacy policies.

It should be noted that the disclosure of the DPO and Data Sharing privacy policy core requirements is mandated only when relevant. However, we chose to include the DPO requirement as Article 29 Data Protection Working Party encourages

the appointment of a DPO even when not mandated (Data Protection Working Party, 2017). Additionally, given the widespread use of third-party content on websites that might transfer user data, we also considered the Data Sharing requirement (Libert et al., 2018).

Furthermore, as to the object of investigation in this study, i.e., the privacy policy, it is worth noting that a privacy policy does not necessarily reflect the actual data processing activities of the data processing organization in question. Based on Article 12 of the GDPR—which states that data processing entities should take appropriate measures to provide any information related to the processing of personal data to the data subject in question in a "concise, transparent, intelligible and easily accessible form, using clear and plain language" (GDPR, 2016)—this study assumes that privacy policies are generally used for that end. Nevertheless, to identify the actual data processing activities, one should explore them in reality rather than their—possibly distorted—reflection, i.e., privacy policies. Future research should explore these actual data processing practices for a more nuanced understanding of GDPR compliance.

Another limitation is that the collected set of privacy policies is limited to what could be collected through web scraping. For that reason, it might have occurred that privacy policies were not collected because they did not meet the scraper's standard. However, it is worth noting the possibility that organizations' privacy policies were not collected because they did not exist.

Furthermore, this study is limited to privacy policies written in English based on practical grounds related to our NLP approach and to reduce the complexity arising from differing interpretations of the GDPR across various languages (Dexe et al., 2022). This limitation might result in an unrepresentative view of the GDPR compliance promises of data processing companies in the EU, as not all organizations target an English-speaking audience. Consequently, a significant proportion of companies in the EU may be excluded from this analysis. Therefore, future studies should aim to replicate these results by considering privacy policies expressed in local languages to provide a more comprehensive understanding of GDPR compliance across the EU.

Moreover, the GDPR applies to all data processing entities that target EU subjects, irrespective of where the processing occurs. Nevertheless, this research is limited to organizations headquartered in the EU as these organizations must—and are therefore more likely to do so, e.g., to avoid repercussions—comply with the GDPR. This approach resulted in a data set that is highly GDPR-relevant and, therefore, suitable for our analysis.

Lastly, further research is necessary to explore the observations that suggest a potential relationship between a country's level of maturity concerning data protection legislation, its history of privacy legislation, and its level of digitization, with the disclosure of GDPR privacy policy requirements in the privacy policies of companies headquartered in that country. These areas warrant further study to enrich our understanding of GDPR compliance and its many influencing factors.

## Supplementary materials

The data set, containing organizational information and GDPR classification (based on the corresponding privacy policy) of 8,614 organizations in the EU, the full output of the analysis, and all scripts regarding preprocessing, scraping, and the analysis, is made available at the following repository: https://aberkane.github.io/GDPR-privacy-policies/.

## References

Aberkane, A., Vanden Broucke, S., & Poels, G. (2022). Investigating organizational factors associated with GDPR noncompliance using privacy policies: A machine learning approach. In *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)* (pp. 107–113). IEEE. https://doi.org/10.1109/TPS-ISA56441.2022.00023

Aberkane, A., Poels, G., & Vanden Broucke, S. (2021). Exploring automated GDPR-compliance in requirements engineering: A systematic mapping study. *IEEE Access*, 9, 66542–66559. https://doi.org/10.1109/ACCESS.2021.3076921

Al-Abdullah, M., Alsmadi, I., AlAbdullah, R., & Farkas, B. (2020). Designing privacy-friendly data repositories: A framework for a blockchain that follows the GDPR. *Digital Policy, Regulation and Governance*, 22(5/6), 389–411.

Almeida Teixeira, G., Mira da Silva, M., & Pereira, R. (2019). The critical success factors of GDPR implementation: A systematic literature review. *Digital Policy, Regulation and Governance*, 21(4), 402–418.

Amaral, O., Abualhaija, S., Torre, D., Sabetzadeh, M., & Briand, L. C. (2022). AI-enabled automation for completeness checking of privacy policies. *IEEE Transactions on Software Engineering*, 48(11), 4647-4674. https://doi.org/10.1109/TSE.2021.3124332

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., & Mayer, J. (2021). Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021* (WWW '21) (pp. 2165–2176). Association for Computing Machinery. https://doi.org/10.1145/3442381.3450048

Barrett, C. (2020). Emerging trends from the first year of EU GDPR enforcement. *Scitech Lawyer, 16*(3), 22-25,35. Retrieved from https://www.proquest.com/scholarly-journals/emerging-trends-first-year-eu-gdpr-enforcement/docview/2369804193/se-2

Bird, S., Klein, E., Loper, E. (2009). *Natural language processing with Python: analyzing text with the Natural Language Toolkit* (1st ed.). O'Reilly Media, Inc.

Bureau van Dijk. (2021). *Orbis database*. Retrieved September 22, 2021, from https://orbis.bvdinfo.com

Bygrave, L. A. (1998). Data protection reforms in Scandinavia. *Privacy Law & Policy Reporter*, 5, 9–12. Retrieved from https://20.austlii.edu.au/cgi-bin/viewdoc/au/journals/PrivLawPRpr/1998/34.html

CMS Law. (2024). GDPR enforcement tracker - list of GDPR fines. Retrieved July 9, 2024, from https://www.enforcementtracker.com/?insights

Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H. W., Palka, P., Sartor, G., & Torroni, P. (2018). Automated processing of privacy policies under the EU general data protection regulation. In *Legal knowledge and information systems,* 313, 51–60. https://doi.org/10.3233/978-1-61499-935-5-51

Daigle, B., & Khan, M. (2020). The EU general data protection regulation: An analysis of enforcement trends by EU data protection authorities. *Journal of International Commerce and Economics*, 1–38.

Data Protection Working Party. (2017). Guidelines on data protection officers ('DPOs').

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2019, October). We value your privacy... now take some cookies. *Informatik Spektrum*, 42(5), 345–346. https://doi.org/10.1007/s00287-019-01201-1

Dexe, J., Franke, U., Söderlund, K., van Berkel, N., Jensen, R. H., Lepinkäinen, N., & Vaiste, J. (2022). Explaining automated decision-making: A multinational study of the GDPR right to meaningful information. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 47(3), 669-697.

EDPB. (2023, May). 1.2 billion euro fine for facebook as a result of EDPB binding decision. Retrieved June 23, 2023, from https://www.edpb.europa.eu/news/news/2023/12-billion-euro-fine-facebook-result-edpb-binding-decision_en

European Commission (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119/1.

European Commission. (2022a, July). The digital economy & society index (DESI). Retrieved February 13, 2023, from https://digital-strategy.ec.europa.eu/en/policies/desi

European Commission. (2022b, July). The digital economy & society index (DESI) - Denmark. Retrieved February 13, 2023, from https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2022

Ford, A., Al-Nemrat, A., Ghorashi, S. A., Davidson, J. (2023). The impact of GDPR infringement fines on the market value of firms. *Information & Computer Security*, 31(1), 51–64.

Francis, J. (2022). The battle for the soul of the GDPR: Clashing decisions of supervisory authorities highlight potential limits of procedural data protection. *Minnesota Law Review: Headnotes*, 107, 167-196.

Frantar, M., Gajšek, M. (2023). Better late than never: Slovenia last EU member state to adopt GDPR implementing act. Retrieved February 13, 2023, from https://www.schoenherr.eu/content/better-late-than-never-slovenia-last-eu-member-state-to-adopt-GDPR-implementing-act/

Freitas, M. C., & Mira da Silva, M. (2018). GDPR compliance in SMEs: There is much to be done. *Journal of Information Systems Engineering & Management*, 3(4), 1-7. https://doi.org/10.20897/jisem/3941

Ghonyan, L. (2017, June 29). Advantages and disadvantages of going public and becoming a listed company. SSRN. https://doi.org/10.2139/ssrn.2995271

Gunnigle, P. & McGuire, D. (2001). Why Ireland? A qualitative review of the factors influencing the location of us multinationals in Ireland with particular reference to the impact of labour issues. *Economic and Social Review*, 32 (1), 43–68.

Hamdani, R. E., Mustapha, M., Amariles, D. R., Troussel, A., Meeùs, S., & Krasnashchok, K. (2021). A combined rule-based and machine learning approach for automated GDPR compliance checking. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 40–49). Association for Computing Machinery. https://doi.org/10.1145/3462757.3466081

Herlin-Karnell, E. (2020). EU data protection rules and the lack of compliance in Sweden. Nordic Journal of European Law, 3(2), 95–103. https://doi.org/10.36969/njel.v3i2.22395

Kapoor, K., Renaud, K., & Archibald, J. (2018, April 5). Preparing for GDPR: Helping EU SMEs to manage data breaches. In *Symposium on Digital Behaviour Intervention for Cyber Security* (pp. 13–20). Society for the Study of Artificial Intelligence and Simulation for Behaviour (AISB).

Karjoth, G., & Schunter, M. (2002). A privacy policy model for enterprises. In *Proceedings of the 15th IEEE Computer Security Foundations Workshop* (pp. 271-281). IEEE. https://doi.org/10.1109/CSFW.2002.1021821

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017, January). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(17), 1-5.

Levis, J., Fischer, P. (2021). GDPR 'glasnost': The Spanish AEPD raises the transparency bar and sanctions two banks. *European Data Protection Law Review* (EDPL), 7, 238.

Libert, T., Graves, L., Nielsen, R. (2018). Changes in third-party content on European news websites after GDPR. *Reuters Institute for the Study of Journalism Reports: Factsheet.*

Liepina, R., Contissa, G., Drazewski, K., Lagioia, F., Lippi, M., Micklitz, H. W., Palka, P., Sartor, G., & Torroni, P. (2019). GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism. In *3rd Workshop on Automated Semantic Analysis of Information in Legal Texts* (Vol. 2385).

Linden, T., Khandelwal, R., Harkous, H., & Fawaz, K. (2020). The privacy policy landscape after the GDPR. In *Proceedings on Privacy Enhancing Technologies* 2020 (pp. 47–64). https://doi.org/10.2478/popets-2020-0004

Liu, S., Zhao, B., Guo, R., Meng, G., Zhang, F., & Zhang, M. (2021). Have you been properly notified? Automatic compliance analysis of privacy policy text with GDPR Article 13. In *Proceedings of the Web Conference 2021 (WWW '21)* (pp. 2154–2164). Association for Computing Machinery. https://doi.org/10.1145/3442381.3450022

Motzfeldt, H.M., Næsborg-Andersen, A. (2020). The Right to Be Forgotten in Denmark. In: Werro, F. (eds) *The Right To Be Forgotten. Ius Comparatum - Global Studies in Comparative Law*, vol 40. Springer, Cham. https://doi.org/10.1007/978-3-030-33512-0_4

Müller, N. M., Kowatsch, D., Debus, P., Mirdita, D., & Böttinger, K. (2019). On GDPR compliance of companies' privacy policies. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings* (pp. 151-159). Springer International Publishing.

Ou-Yang, L. (2013). Newspaper: Article scraping & curation. Retrieved from https://github.com/codelucas/newspaper

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhoger, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011, November). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.

Presthus, W., Sønslien, K. F. (2021). An analysis of violations and sanctions following the GDPR. *International Journal of Information Systems and Project Management*, 9 (1), 38–53.

Al Rahat, T., Long, M., & Tian, Y. (2022). Is your policy compliant? A deep learning-based empirical study of privacy policies' compliance with GDPR. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society (WPES'22)* (pp. 89–102). Association for Computing Machinery. https://doi.org/10.1145/3559613.3563195

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J. T., Liu, F., McDonald, A., Norton, T. B., Ramanath, R., Russell, N. C., Sadeh, N., Schaub, F. (2015). Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30 (1), 39–88.

Riccardi, J. L. (1983). The German federal data protection act of 1977: Protecting the right to privacy. *Boston College International and Comparative Law Review*, 6, 243.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503-520. https://doi.org/10.1108/00220410410560582

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 92-96). https://doi.org/10.25080/Majora-92bf1922-011

Serrado, J., Pereira, R. F., Mira da Silva, M., & Scalabrin Bianchi, I. (2020). Information security frameworks for assisting GDPR compliance in the banking industry. *Digital Policy, Regulation and Governance*, 22(3), 227-244.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Strzelecki, A., & Rizun, M. (2020). Consumers' security and trust for online shopping after GDPR: Examples from Poland and Ukraine. *Digital Policy, Regulation and Governance*, 22(4), 289-305.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018). PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics (IWSPA '18)* (pp. 15-21). Association for Computing Machinery. https://doi.org/10.1145/3180445.3180447

Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2018). EU general data protection regulation: Changes and implications for personal data collecting companies. *Computer Law & Security Review*, 34(1), 134-153. https://doi.org/10.1016/j.clsr.2017.05.015

Vilas, M. (2020). Googlesearch. Retrieved from https://github.com/MarioVilas/googlesearch

Vrabec, H. U. (2020). Slovenia: Introduction to the most recent public draft of the GDPR implementing law. *European Data Protection Law Review (EDPL)*, 6, 121.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14)* (Article 38, pp. 1–10). Association for Computing Machinery. https://doi.org/10.1145/2601248.2601268

Zaeem, R. N., Anya, S., Issa, A., Nimergood, J., Rogers, I., Shah, V., Srivastava, A., & Barber, K. S. (2020). PrivacyCheck v2: A tool that recaps privacy policies for you. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)* (pp. 3441–3444). Association for Computing Machinery. https://doi.org/10.1145/3340531.3417469

Zaeem, R. N., & Barber, K. S. (2021). Comparing privacy policies of government agencies and companies: A study using machine-learning-based privacy policy analysis tools. In Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART) (Vol. 2, pp. 29–40). SciTePress. https://doi.org/10.5220/0010180700290040

Zell, A. M. (2014). Data protection in the federal republic of Germany and the European Union: An unequal playing field. *German Law Journal*, 15 (3), 461–494.

## Biographical notes

Abdel-Jaouad Aberkane received the B.Sc. degree in computing science from the University of Amsterdam (Amsterdam, the Netherlands) in 2016. Subsequently, he received the M.Sc. degree in information science from Utrecht University (Utrecht, the Netherlands) in 2018. He is currently pursuing the Ph.D. degree with the Business Informatics Research Group on automated GDPR-compliance in requirements engineering at Ghent University (Ghent, Belgium). His research interests include natural language processing, requirements engineering, and the General Data Protection Regulation.
ORCID: 0000-0002-4557-0715

Seppe Vanden Broucke is currently working as an associate professor at the Department of Business Informatics and Operations Management at UGent (Belgium) and is a guest lecturer at KU Leuven (Belgium). Seppe's research interests include machine learning, business analytics, and process mining. His work has been published in well-known international journals and presented at top conferences. Seppe received his PhD in Applied Economics at KU Leuven, Belgium in 2014. Seppe is co-holder of several research chairs, is the co-academic organizer of Postgraduate in Big Data & Analytics in Business and Management, and has authored several books on machine learning and data related topics.
ORCID: 0000-0002-8781-3906

Geert Poels is Senior Full Professor of Business Informatics at the Faculty of Economics and Business Administration, Ghent University (Ghent, Belgium), where he teaches intermediate and advanced courses on Information Systems, IT Management, Enterprise Architecture, and Service Design. He also teaches in the Master of Enterprise ICT Architecture at IC Institute (Beerzel, Belgium). His research relates to Conceptual Modeling (as research method) and Enterprise Modeling (as research domain) with a focus on business process architecture mapping, ArchiMate, Value Modeling, and NLP-based automated generation of conceptual models out of user requirements documents. He also supervises Ph.D. research on digital marketplaces, cybersecurity and GDPR. As academic service, he was member of the development team of the COBIT 2019 framework for IT governance.
ORCID: 0000-0001-9247-6150