Investigating Organizational Factors Associated with GDPR Noncompliance using Privacy Policies: A Machine Learning Approach

1st Abdel-Jaouad Aberkane
Department of Business Informatics and Operations Management
Ghent University
Ghent, Belgium
0000-0002-4557-0715

2nd Seppe vanden Broucke
Department of Business Informatics and Operations Management
Ghent University
Ghent, Belgium
0000-0002-8781-3906

3rd Geert Poels
Department of Business Informatics and Operations Management
Ghent University
Ghent, Belgium
0000-0001-9247-6150

***Abstract*—The General Data Protection Regulation (GDPR) came into effect in May 2018 to ensure and safeguard data subjects' rights. This enactment profoundly shaped, among other things, data processing organizations' privacy policies to comply with the GDPR's transparency requirements—for compliance with the GDPR is compulsory. Nevertheless, despite the potential goodwill to change, complying with the GDPR can be challenging for some organizations, e.g., small and medium-sized enterprises, due to, for example, a lack of resources. This study explores what factors may correlate with GDPR-compliance practices in organizations by analyzing the corresponding privacy policies. The contribution of this study is twofold. First, we have devised a classification model using machine learning (ML) and natural language processing (NLP) techniques to assess the GDPR-compliance practices promised in privacy policies regarding the GDPR core privacy policy requirement of Purpose. Using this model, we have collected a data set of 8 614 organizations active in the European Union (EU) containing organizational information and GDPR-compliance promises derived from organizations' privacy policies, as made publicly available. Our second contribution is an analysis of the resulting classification to identify organizational factors related to the disclosure of the GDPR core privacy policy requirement of Purpose in organizations' privacy policies.**

***Index Terms*—General Data Protection Regulation, Privacy, Machine Learning, Natural Language Processing**

I. Introduction

The advent of the General Data Protection Regulation (GDPR) resulted in regulations concerning privacy relevant to data processing entities [1]. Consequently, data processing entities shifted their attention to compliance with the GDPR, which is reflected in the increase of transparency on the Internet. More organizational websites have privacy policies and notify users regarding their data processing practices [2]. However, despite this effort, four years after the GDPR was provisionally agreed upon, GDPR-compliance proved to be challenging to many businesses as they, among other things, lack the adequate tools for tracking their regulation and consumer engagement obligations [3]. These challenges may be due to, for example, the lack of resources or expertise (i.e., resource poverty)—which is typically related to small and medium-sized enterprises [4], [5]—to realize the appropriate technical and organizational measures for GDPR-compliance as outlined in the GDPR [1].

More factors may correlate with GDPR-compliance than size and resources alone. For example, the sector in which the data processing entity is active and geographical location may play a role as well: government agencies are better at protecting user data than companies, and privacy policies of European government agencies perform better than their United States counterparts in, e.g., giving users the right to edit or delete their data [6]. However, both sector and location fail to describe security measures toward protecting user data, indicating that, still, steps need to be taken on the road to GDPR-compliance. In fact, research shows that a significant number of GDPR-compliance practices of organizations, as reflected in privacy policies, do not comply with the GDPR [7]–[9].

Centering on the relationship between organizational factors and GDPR-compliance, this study analyzes data processing entities' organizational factors to identify those factors that correlate with GDPR-compliance practices as reflected in privacy policies. In particular, we focus on GDPR-compliance practices with regards to one of the GDPR core privacy policy requirements, namely *Purpose*, as it is one of the central themes of the GDPR and deemed to be generic and easily identifiable in a given privacy policy [9]. Purpose comprises—as described in Article 13 (§1c) of the GDPR—the purposes of the processing for which the personal data are intended, as well as the legal basis for processing [1].

In order to assess the GDPR-compliance practices of organizations, this study takes their privacy policy, as communicated on their website, as a point of reference since it is the most important source of information for the general public concerning the data processing practices of the corresponding data processing entity [10]. In fact, the GDPR propelled the most widespread changes to privacy policies in the last decade [11]. Moreover, after the commencement of the GDPR, privacy policies increased in length and covered categories

of particular importance to the GDPR requirements [12]. Furthermore, websites that use GDPR terminology reflect an increase in information about users' rights and the legal basis of processing [2]. Given that privacy policies are expressed in natural language, this study uses natural language processing (NLP) techniques combined with machine learning (ML) to identify organizational factors that correlate with GDPR-compliance practices.

To our knowledge, no research has identified factors that correlate with GDPR-compliance practices, as communicated in privacy policies published on organizational websites, using ML and NLP. This study aims to fill this gap and act as a stepping stone to guide future research toward a refined attitude in achieving GDPR-compliance. To do so, we adopt an approach consisting of three phases. The first phase uses NLP to train a logistic regression model to classify privacy policies based on one of the GDPR core requirements, Purpose, benefiting from an existing annotated data set of 250 privacy policies containing over 18 300 natural sentences. A privacy policy is classified as positive if a proportion of its sentences disclose the purpose of processing. The calibration of this proportion is explained in detail in Section IV. In the second phase, a data set of organizations active in the European Union (EU) was collected, including the corresponding organizational data such as size, sector, and location. Then, based on the organization's name, the related privacy policy was—if available—scraped and subjected to the classification model of the first phase of this study. This resulted in a new data set containing organizational data and classification results. Finally, this final data set was analyzed to identify organizational factors associated with disclosing the considered GDPR core privacy policy requirement of Purpose in the organization's privacy policy.

The rest of this paper is organized as follows. Section 2 describes the related work. In Section 3, the adopted research approach is presented. Section 4 focuses on the privacy policy classification. Section 5 discusses the data set creation as used in our analysis. The latter, including the corresponding results, are outlined in Section 6. Section 7 describes the discussion and threats to the validity of our approach. Finally, Section 8 concludes our research and provides pointers to future work.

II. Related Work

In the following, a brief overview of existing research related to privacy policies, GDPR, and ML is collected through the (reverse) snowballing approach [13]. This overview reveals that—to the best of our knowledge—no research has been conducted identifying organizational factors that correlate with GDPR-compliance practices, as stated in online privacy policies, while using ML and NLP.

The majority of relevant research centers on evaluating privacy policy completeness based on the GDPR. For example, El Hamdani et al. develop methods to verify compliance of privacy policies to the GDPR by incorporating rule-based approaches and ML [14]. In a similar vein, Liu et al. present an approach to automatically analyze privacy policy contents and identify violations against Article 13 of the GDPR using ML and rule-based analysis [15]. Amaral et al. propose an AI-enabled approach for completeness checking of privacy policies according to the GDPR, using 234 privacy policies from the fund industry to evaluate their approach [16]. Liepin et al. work toward a methodology for annotating post-GDPR privacy policies to identify and assess their compliance with the regulation using legal analysis, ML, and NLP [17]. Müller et al. introduce a data set of annotated privacy policies; each sentence snippet was labeled concerning its compliance with five GDPR core privacy policy requirements [9]. Furthermore, the authors evaluate the validity of the data set by using NLP algorithms in combination with supervised learning techniques.

The literature discussed above is restricted to GDPR-centered studies. However, it is worth noting that ML-based privacy policy completeness has been researched before the advent of the GDPR. Costante et al., for example, present a system to assess, using ML, the completeness of a privacy policy based on, among other sources, the predecessor of the GDPR, i.e., EU 95/46/EC [18].

Despite the developments related to GDPR-compliance, there is still considerable work to be done for data processing organizations. Al Rahat, Le, and Tian, conclude, for example, after creating an annotated privacy policy data set upon which they apply a convolutional neural network based model, that even after the GDPR went into effect, the vast majority of websites (97%) still failed to comply with at least one requirement of the GDPR [7]. Furthermore, Contissa et al. conducted an experimental study using ML to evaluate privacy policies under the GDPR, concluding that none of the analyzed privacy policies gets close to the standards of the GDPR [8]. Along the same lines, Zaeem and Barber conclude that the GDPR has made progress in protecting user data, but, "more progress is necessary", after investigating the effect of the GDPR on privacy policies using ML [19]. Finally, a recent study by Zaeem and Barber shows a slight overlap with the research goal of this study. The authors use ML tools to explore *if* and *how* companies and government agencies differ in their privacy policies, demonstrating that European government agencies' privacy policies perform better than their United States peers concerning several GDPR requirements [6].

Our study, however, identifies organizational factors (e.g., size, geographical location) that correlate with GDPR-compliance practices using a data set of 8 614 data processing entities' privacy policies. Moreover, this data set is not restricted to companies and government agencies. In fact, in addition to other factors, we distinguish between 21 economic activities as per NACE Rev. 2 [20], as discussed further in Section III.

III. Research Approach

The approach used in this study consists of three stages, as depicted in Fig. 1. In the first stage, presented in more detail in Section IV, we benefit from an existing annotated

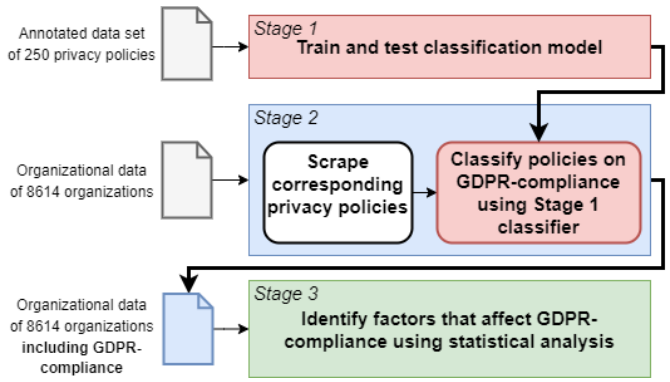

Fig. 1. High level overview of the three-staged approach of this study.

data set of 250 privacy policies—comprising over 18 300 natural sentences—labeled according to five GDPR core privacy policy requirements [9]. Next, using NLP, linguistic features were identified to train a classification model based on logistic regression to classify whether or not privacy policies disclose the purpose of processing, thus meeting the GDPR core requirement of Purpose.

In the second stage, detailed in Section V, we collected organizational data of 168 824 companies located in the EU from the Orbis database, provided by Bureau van Dijk [21]. Then, we scraped—if possible—the related privacy policies of the gathered companies. This scraping process resulted in 8 614 privacy policies. Subsequently, utilizing the classification model of stage one, the scraped privacy policies were classified on the GDPR core requirement of Purpose. Finally, the classification output was combined with the initial data set containing organizational factors and used for analysis in stage three. In this last stage, the combined data set—containing both organizational factors as the classification results—was analyzed to identify organizational factors that correlate with the disclosure of the GDPR core requirement of Purpose in the organization's privacy policy, using the organizational factors as predictors and Purpose as a target value. Section VI expands upon this stage.

IV. Privacy Policy Classification

In this study, we make use of the data set by Müller et al. containing 18 397 labeled sentences—making up 250 privacy policies—labeled according to five GDPR core requirements [9]. The data set was obtained by automatically crawling and storing privacy policies. After that, the collected privacy policies' sentences were manually labeled into one or more of the following five classes (i.e., GDPR privacy policy core requirements): *Data Protection Officer*, *Purpose*, *Acquired Data*, *Data Sharing*, and *Rights*. The data set contains 18 397 sentences, of which 971 are labeled as relevant to the class of our interest, i.e., Purpose. Moreover, the Purpose class is covered in 88.8% of the 250 privacy policies. A sentence is classified as compliant with the Purpose class if the purpose for processing is disclosed.

TABLE I
Initial sentence classification performance.

GDPR Class	Precision	Recall	F1	AUC-score	N-gram configuration
Purpose	0.45	0.76	0.56	0.956	(1,3)

Our first step toward classification encompassed preprocessing the data set using Python's *Natural Language Toolkit* (NLTK) [22] library. This step consisted of executing the standard NLP pipeline: tokenization, removing punctuation, removing digits, removing stopwords, and stemming. Subsequently, the preprocessed sentences were vectorized using Term Frequency - Inverse Document Frequency (TFIDF) [23]. We used TFIDF scores of unigrams, bigrams, and trigrams—which are instances of n-grams (i.e., sequences of tokens of length n)—as features for the classification model.

To address the class imbalance (about 95% of the sentences did not belong to the Purpose class), we conducted a stratified split (using the *scikit-learn* library [24]) of the data into a training and test set, followed by an oversampling of the minority class in the training set. Then, we optimized the n-gram length through cross-validation to construct our classification model. For the classification, we opted for logistic regression, a well-known, interpretable, and suitable technique for binary supervised classification tasks.

Table I presents the performance of the trained model on sentence level. Since our focus lies in assessing the GDPR-compliance practices surrounding Purpose as disclosed in privacy policies, the following questions arise when interpreting the classification results: *When does a privacy policy as a whole meet the Purpose requirement? Is the presence of a single positively labeled sentence sufficient to classify a privacy policy as compliant with the Purpose class?* To increase the confidence of our predictions, we address the following problem: *What number of positive sentences are needed to classify, with a desired degree of precision, a new document as compliant with the requirement at issue?* Our approach consists of setting a threshold employing the inverse cumulative distribution function of the binomial distribution.

Given 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, E = TP + FP + TN + FN elements—$\hat{P}$ describes the elements predicted as positive, i.e., $\hat{P}$ = 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:

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

Then, the cumulative distribution function is:

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

We are interested in the probability that the number of TP exceeds a given value, thus we consider the inverse cumulative distribution:

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

We now find the highest value of k that keeps the inverse cumulative distribution above a given desired probability Z:

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

Given a new set of elements (e.g., sentences in a privacy policy) with E_n elements of which $\hat{P}_n$ are predicted as positive by the model, we consider the document to be positive if the threshold, $Q = \frac{k'}{E}$ is met, i.e., $\frac{\hat{P}}{E_n} \geq Q$.

We have calibrated the threshold Q for the Purpose class, using the test set size (E_{test} equal to 1 840) while setting the desired probability (i.e., Z) at 90%. The calibration reveals that for a given new set of elements (e.g., sentences of a privacy policy), the following threshold $\frac{\hat{P}}{E_n} \geq 0.035$ must be met to achieve a precision of at least 0.909, realizing a substantial increase compared with the initial precision of 0.45.

V. Data Set Creation

First, the Orbis database—containing information of close to 400 million companies and entities across the globe—was utilized to collect a random sample of data comprising information of 168 824 companies located in Europe [21]. This data, coined organizational data in this study, included the following information: the **company name**, **quoted** (describing whether the company was publicly listed), the **country ISO code** indicating the location of the company, **NACE code** describing the industrial classification 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, lastly, the **size classification** of the organization. The company name was used, as explained in Section V, for scraping privacy policies but was later omitted from the data set. In addition, the last available year was used as a filter to retain only data relevant to the GDPR (i.e., data from 2018 and onward) and was not used as a predictor in the final analysis. Also, as mentioned before, we only included companies located in the EU as they are more likely to comply with the GDPR since they are obliged to do so, hence collecting a GDPR-relevant data set.

Regarding the size classification, we diverted from the classification used by Orbis (i.e., *small companies*, *medium-sized companies*, *large companies*, and *very large companies*), by grouping the former two under "small and medium-sized enterprises", and by grouping the latter two under "large enterprises". Large enterprises are characterized by meeting at least one of the following criteria: an operating revenue equal to or more than 10 million euro, possessing total assets equal to or more than 20 million euro, and having 150 employees or more. Small and medium-sized enterprises are those enterprises that do not meet these criteria. The final set of considered organizational factors is presented in Table II.

TABLE II
Final set of organizational factors considered for analysis.

Name	Description
Quoted	Boolean value indicating whether the company is listed or not.
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".
Operating revenue	Operating revenue as reported in the last available year.
Number of employees	Number of employees as reported in the last available year.
Size classification	Boolean value describing the size of the company (i.e., small and medium-sized or large enterprise).

After collecting the data, a scraper was developed to collect the privacy policies of the companies considered. The scraper identifies the relevant privacy policies using Google search results, utilizing Python's *googlesearch* [25], *Newspaper* [26], and NLTK libraries [22]. For each company, a Google query was composed incorporating the company name merged with the words "privacy policy". Then, the first three results were assessed for their relevancy, only selecting URLs that contained the company name in the domain name, thus avoiding third-party websites that mention the company name in their URL. Moreover, the URL must contain the words *privacy* or *policy* to avoid scraping irrelevant information. After identifying the relevant URL, the content of the corresponding web page was scrutinized as it had to be in English and of considerable length. If these requirements were met, the privacy policy was scraped. Following the scraping procedure, a random sample of the scraped policies was manually examined, ensuring the scraper's effectiveness and correctness.

VI. Analysis

Before conducting the analysis with our descriptive logistic regression model, we converted the data types of the predictors (i.e., organizational factors) and the target value (i.e., compliance with the Purpose requirement) to a categorical and numerical representation. Subsequently, the categorical data was encoded, while numerical data was scaled to suit the analysis.

The next step consisted of parameter optimization after splitting the data set randomly into a training set and test set: the logistic regression parameters were optimized based on the training set, after which the performance was evaluated

using the test set. The optimization resulted in a classification accuracy of 0.66, using L1 regularization, to acquire a parsimonious model, and an alpha value (i.e., the weight multiplying the L1 penalty term) of 2.101. Thereafter, the whole data set was retrained using the optimized parameters. The complete output of the analysis—at the 0.05 significance level—can be found at the following repository: https://aberkane.github.io/Privacy-Policies-GDPR-compliance.

A. Results

Table III presents the statistically significant predictors and corresponding coefficients for the GDPR requirement of Purpose. First, the predictor *quoted* positively correlates with the target variable of Purpose, meaning, a publicly listed organization is more likely to disclose its data processing practices concerning Purpose in its privacy policy than an organization that it is not listed. The same is true for the following countries (*country ISO code*) that host data processing entities' headquarters: Belgium, Denmark, Spain, France, Greece, Ireland, the Netherlands, and Sweden. On the other hand, Germany and Lithuania negatively correlate with complying with the Purpose requirement. Regarding the industry classification (*NACE rev. 2*), "financial and insurance activities" and "information and communication" contribute positively to complying with the Purpose requirement. Contrastingly, "agriculture, forestry and fishing" and "manufacturing" reveal a negative relationship toward compliance with Purpose. Similarly, predictor "small and medium-sized enterprises" (*size classification*) negatively correlates with compliance with the Purpose requirement.

TABLE III
SIGNIFICANT PREDICTORS AND CORRESPONDING COEFFICIENTS FOR THE TARGET VALUE OF PURPOSE.

Predictor	**P-value**	**Coefficient**
Quoted	0.048400	0.324 544
BE (Country ISO code)	0.000172	0.507 113
DE (Country ISO code)	$3.330\,460 \times 10^{-13}$	−0.543 350
DK (Country ISO code)	0.003432	0.578 215
ES (Country ISO code)	0.011291	0.226 202
FR (Country ISO code)	$1.989\,637 \times 10^{-5}$	0.355 348
GR (Country ISO code)	$7.855\,161 \times 10^{-4}$	0.655 847
IE (Country ISO code)	$3.773\,525 \times 10^{-7}$	0.822 251
LT (Country ISO code)	$9.732\,925 \times 10^{-4}$	−0.694 304
NL (Country ISO code)	$1.577\,728 \times 10^{-5}$	0.609 118
SE (Country ISO code)	$9.974\,404 \times 10^{-9}$	0.646 488
Agriculture, forestry and fishing (NACE Rev. 2 code)	0.014787	−0.677 507
Financial and insurance activities (NACE Rev. 2 code)	$1.327\,102 \times 10^{-5}$	0.421 090
Information and communication (NACE Rev. 2 code)	0.022121	0.225 640
Manufacturing (NACE Rev. 2 code)	0.028305	−0.149 008
Small and medium-sized enterprises (Size classification)	0.011749	−0.124 968

VII. DISCUSSION

This section discusses the statistical analysis results and describes the threats to the validity of this study. A complete overview of the results, including all considered variables and their subcategories, can be found in the repository mentioned in Section VI.

Concerning the **country ISO code**, or the country in which the data processing organization is headquartered, we observe that a minority of the countries (10 of the 27 countries of the EU) correlate with the GDPR requirement of Purpose. Focusing on these significant predictors, we notice that the vast majority—eight countries—positively correlate with the target value of Purpose. Regarding the industry classification as per **NACE rev. 2**, we observe that, from a total of 21, only four different sectors correlate with target value Purpose: two sectors positively correlate with GDPR-compliance, whereas two had a negative correlation. The remaining 17 sectors did not significantly correlate with the GDPR core requirement of Purpose. Centering on the predictor **quoted** (i.e., indicating whether a company is listed or not), we observe that it positively correlates with GDPR-compliance practices. On the other hand, being categorized as a small or medium-sized enterprise (**size classification**) negatively correlates with GDPR-compliance. Lastly, the predictors of **operating revenue** and **number of employees** were not proven to be significant.

Concretely, the reported findings can steer future research and practices as they present factors that correlate with GDPR-compliance as reflected in organizations' privacy policies. In particular, the factors that negatively correlate with GDPR-compliance practices can be taken as a guideline to address the problem of noncompliance. Consequently, these findings can lead to a more refined attitude, in both research and practice, toward addressing GDPR-compliance.

A. Threats to Validity

First, this study relies on one GDPR core privacy policy requirement, whereas the GDPR itself is more comprehensive. We opted for this approach to limit the scope and focus on, as argued before, the core requirement of Purpose that takes a central role in the GDPR and is considered to be readily deducible from privacy policies. Furthermore, as to the latter, privacy policies do not necessarily reflect the actual data processing activities of the corresponding data processing entity. However, in this study, we assume that they should reasonably reflect the data processing practices as per GDPR Article 12: 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 [1].

In addition, this study is limited to privacy policies identifiable by our scraper. Therefore, it may be possible that our scraper could not identify or collect the privacy policies of some organizations due to technical reasons despite it being published on their website, which is a limitation of this study. Furthermore, it may also be possible that organizations did not publish a privacy policy at all on their website, despite being data processing entities. However, this falls out of the scope of this study.

Another limitation of this study is that only privacy policies expressed in the English language were considered. This might provide a skewed view of the GDPR-compliance practices of the data processing companies in the EU, as organizations might prefer, especially if they are not active in an international context, to express their privacy policy in their local language. Nevertheless, it was decided to consider only privacy policies written in the English language for practical reasons related to the adopted natural language processing approach.

Finally, the data set in Section V is limited to organizations headquartered in the EU only. However, this is not necessarily the only geographical location to which the GDPR is relevant. In fact, the GDPR is relevant to all data processing entities that process data of EU data subjects. Having said that, since organizations in Europe are more likely to comply as they are obliged to do so, we have opted to limit the organizations to those located in the EU, thus creating a more GDPR-relevant data set for the conducted analysis.

VIII. Conclusion

This study describes an approach to identify organizational factors that correlate with GDPR-compliance practices in organizations. In particular, this study focuses on one of the GDPR core privacy policy requirements, namely, Purpose. The first stage of the approach involves devising a classification model using ML and NLP, using an annotated data set consisting of over 18 300 sentence snippets, achieving a precision of at least 0.909 in classifying privacy policies on their disclosure of the purpose of data processing. In the second stage, we departed from a data set containing organizational information of 8 614 companies, scraped the related privacy policies, and, finally, classified them using the trained classification model of stage one. Thus, assembling a data set containing companies' organizational factors and GDPR-compliance practices surrounding the Purpose requirement as reflected in the corresponding privacy policy. Eventually, the last stage consisted of subjecting the collected data set to analysis toward identifying organizational factors that correlate with the disclosure of the core requirement of Purpose in the related privacy policy.

The findings of this study shed a nuanced light on the problem of GDPR noncompliance. In particular, we conclude that being a publicly listed company positively correlates with GDPR-compliance practices regarding Purpose. On the other hand, being a small or medium-sized enterprise negatively correlates with complying with the Purpose requirement. Furthermore, the results show that most of the considered geographical locations and industry classifications (i.e., sectors) do not correlate with compliance with the Purpose requirement. These findings, both predictive and non-predictive factors, provide handles to researchers and entities aiming to address GDPR-compliance, albeit theoretically or practically (e.g., tooling), by pinpointing factors that should, at least, be regarded.

References

[1] The European Parliament and the Council of European Union, "REGULATION (EU) 2016/679," *Off. J. Eur. Union*, pp. 1–2, Apr 2016.

[2] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, "We value your privacy ... now take some cookies," *Informatik Spektrum*, vol. 42, no. 5, pp. 345–346, Oct 2019.

[3] Thomson Reuters, "GDPR +1 year : Business struggles with data privacy regulations increasing," May 2019, accessed on Dec. 15, 2021. [Online]. Available: http://ask.legalsolutions.thomsonreuters.info/GDPR1YearBusinessStrugglesReport

[4] M. C. Freitas and M. Mira da Silva, "GDPR compliance in SMEs: There is much to be done," *J. Inf. Syst. Eng. & Manage.*, vol. 3, no. 4, p. 30, 2018.

[5] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "EU general data protection regulation: Changes and implications for personal data collecting companies," *Comput. Law & Sec. Review*, vol. 34, no. 1, pp. 134–153, 2018.

[6] R. N. Zaeem and K. Barber, "Comparing privacy policies of government agencies and companies: A study using machine-learning-based privacy policy analysis tools," in *Proc. 13th Int. Conf. Agents Artif. Intell. - Volume 2: ICAART,*, INSTICC. SciTePress, 2021, pp. 29–40.

[7] T. A. Rahat, T. Le, and Y. Tian, "Automated detection of gdpr disclosure requirements in privacy policies using deep active learning," *arXiv preprint arXiv:2111.04224*, 2021.

[8] G. Contissa, K. Docter, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Palka, G. Sartor, and P. Torroni, "Automated processing of privacy policies under the EU general data protection regulation," in *31st Int. Conf. Legal Knowl. Inf. Syst.*, vol. 313. IOS Press, 2018, pp. 51–60.

[9] N. M. Müller, D. Kowatsch, P. Debus, D. Mirdita, and K. Böttinger, "On GDPR compliance of companies' privacy policies," in *Int. Conf. Text, Speech, Dialogue*. Springer, 2019, pp. 151–159.

[10] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. McDonald, T. B. Norton, R. Ramanath, N. C. Russell, N. Sadeh, and F. Schaub, "Disagreeable privacy policies: Mismatches between meaning and users' understanding," *Berkeley Technol. Law J.*, vol. 30, no. 1, pp. 39–88, 2015.

[11] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, "Privacy policies over time: Curation and analysis of a million-document dataset," in *Proc. Web Conf. 2021*, 2021, pp. 2165–2176.

[12] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the gdpr," *Proc. Privacy Enhancing Technol.*, vol. 1, pp. 47–64, 2020.

[13] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quart.*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[14] R. E. Hamdani, M. Mustapha, D. R. Amariles, A. Troussel, S. Meeùs, and K. Krasnashchok, "A combined rule-based and machine learning approach for automated GDPR compliance checking," in *Proc. 18th Int. Conf. Artif. Intell. Law*. ACM, 2021, pp. 40–49.

[15] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang, "Have you been properly notified? automatic compliance analysis of privacy policy text with GDPR article 13," in *Proc. Web Conf.* New York, NY, USA: ACM, 2021, p. 2154–2164.

[16] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, and L. Briand, "AI-enabled automation for completeness checking of privacy policies," *IEEE Trans. Softw. Eng.*, Nov 2021.

[17] R. Liepina, G. Contissa, K. Drazewski, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Palka, G. Sartor, and P. Torroni, "Gdpr privacy policies in claudette: Challenges of omission, context and multilingualism," in *CEUR Workshop Proc.*, vol. 2385. CEUR-WS, 2019.

[18] E. Costante, Y. Sun, M. Petković, and J. Den Hartog, "A machine learning solution to assess privacy policy completeness: (short paper)," in *Proc. ACM Workshop Privacy Electron. Soc.* ACM, 2012, pp. 91–96.

[19] R. N. Zaeem and K. S. Barber, "The effect of the gdpr on privacy policies: Recent progress and future promise," *ACM Trans. Manage. Inf. Syst.*, vol. 12, no. 1, Dec 2020.

[20] H. Carré, "Statistical classification of economic activities in the european community," *Publications Office of the European Union*, 2008.
[21] Bureau van Dijk, "Orbis database," https://orbis.bvdinfo.com/, accessed on Sept. 22, 2021.
[22] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the Natural Language Toolkit*, 1st ed. O'Reilly Media, Inc., 2009.
[23] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, 2004.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, p. 2825–2830, Nov 2011.
[25] M. Vilas, "googlesearch," *Python Library. Retrieved*, 2020. [Online]. Available: https://github.com/MarioVilas/googlesearch
[26] L. Ou-Yang, "Newspaper: Article scraping & curation," *Python Library. Retrieved*, 2013. [Online]. Available: https://github.com/codelucas/newspaper