COMPUTER RESEARCH AND MODELING
2024 VOL. 16 NO. 7 P. 1667–1685
DOI: 10.20537/2076-7633-2024-16-7-1667-1685

Ки&М

SPECIAL ISSUE

UDC: 004.912

NLP-based automated compliance checking of data processing agreements against General Data Protection Regulation

O. N. Okonicha[a], A. Sadovykh[b]

Innopolis University,
1 Universitetskaya st., Innopolis, 420500, Russia

E-mail: [a] o.okonicha@innopolis.university, [b] a.sadovykh@innopolis.ru

Received 28.10.2024, after completion — 13.11.2024
Accepted for publication 25.11.2024

As it stands in the contemporary world, compliance with regulations concerning data protection such as GDPR is central to organizations. Another important issue analysis identified is the fact that compliance is hampered by the fact that legal documents are often complex and that regulations are ever changing. This paper aims to describe the ways in which NLP aids in keeping GDPR compliance effortless through automated scanning for compliance, evaluating privacy policies, and increasing the level of transparency. The work does not only limit to exploring the application of NLP for dealing with the privacy policies and facilitate better understanding of the third-party data sharing but also proceed to perform the preliminary studies to evaluate the difference of several NLP models. They implement and execute the models to distinguish the one that performs the best based on the efficiency and speed at which it automates the process of compliance verification and analyzing the privacy policy. Moreover, some of the topics discussed in the research deal with the possibility of using automatic tools and data analysis to GDPR, for instance, generation of the machine readable models that assist in evaluation of compliance. Among the evaluated models from our studies, SBERT performed best at the policy level with an accuracy of 0.57, precision of 0.78, recall of 0.83, and F1-score of 0.80. BERT showed the highest performance at the sentence level, achieving an accuracy of 0.63, precision of 0.70, recall of 0.50, and F1-score of 0.55. Therefore, this paper emphasizes the importance of NLP to help organizations overcome the difficulties of GDPR compliance, create a roadmap to a more client-oriented data protection regime. In this regard, by comparing preliminary studies done in the test and showing the performance of the better model, it helps enhance the measures taken in compliance and fosters the defense of individual rights in the cyberspace.

Keywords: compliance audit, NLP, DPA, GDPR, privacy, SBERT, BERT, GPT

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 7, pp. 1667–1685.

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2024 Т. 16 № 7 С. 1667–1685
DOI: 10.20537/2076-7633-2024-16-7-1667-1685

Ки&М

СПЕЦИАЛЬНЫЙ ВЫПУСК

УДК: 004.912

Автоматизированная проверка соответствия соглашений об обработке данных регламенту по защите данных

О. Н. Оконича[a], А. Садовых[b]

Университет Иннополис,
Россия, 420500, г. Иннополис, ул. Университетская, д. 1

E-mail: [a] o.okonicha@innopolis.university, [b] a.sadovykh@innopolis.ru

Получено 28.10.2024, после доработки — 13.11.2024
Принято к публикации 25.11.2024

В современном мире соблюдение нормативных требований по защите данных, таких как GDPR, является ключевым для организаций. Другой важной проблемой, выявленной при анализе, является то, что соблюдение осложняется сложностью правовых документов и постоянными изменениями в регулировании. В данной статье описываются способы, с помощью которых NLP (обработка естественного языка) способствует упрощению соблюдения GDPR путем автоматического сканирования на соответствие, оценки политик конфиденциальности и повышения уровня прозрачности. Работа не ограничивается исследованием применения NLP для работы с политиками конфиденциальности и улучшения понимания обмена данными с третьими сторонами, но также проводит предварительные исследования для оценки различий между несколькими моделями NLP. В статье описывается реализация и исполнение моделей для выявления той, которая демонстрирует наилучшую производительность по эффективности и скорости автоматизации процесса проверки соответствия и анализа политики конфиденциальности. Кроме того, в исследовании обсуждаются возможности использования автоматических инструментов и анализа данных для соблюдения GDPR, например, создание машиночитаемых моделей, которые помогают в оценке соответствия. Среди моделей, оцененных в нашем исследовании, SBERT показала лучшие результаты на уровне политики с точностью 0,57, прецизионностью 0,78, полнотой 0,83 и F1-метрикой 0,80. Модель BERT продемонстрировала наивысшую производительность на уровне предложений, достигнув точности 0,63, прецизионности 0,70, полноты 0,50 и F1-метрики 0,55. Таким образом, данная статья подчеркивает важность NLP в помощи организациям преодолеть трудности соблюдения GDPR, создавая дорожную карту к более ориентированному на клиента режиму защиты данных. В этом отношении, сравнивая предварительные исследования и демонстрируя производительность лучших моделей, работа способствует усилению мер по соблюдению и защите прав личности в киберпространстве.

Ключевые слова: аудит соответствия, NLP (обработка естественного языка), DPA (соглашение об обработке данных), GDPR (общий регламент по защите данных), конфиденциальность, SBERT, BERT, GPT

Introduction

The vast amount of personal information disclosed and collected in the contemporary world increases essential issues about confidentiality and data safeguarding. The General Data Protection Regulation (GDPR) is a complex law adopted by the European Union that seeks to safeguard the rights of EU citizens whose data has been collected by organizations by placing strict consequence on organizations using and processing such data [Voigt, von dem Bussche, 2017; Alattas et al., 2022]. Personal data is respected with GDPR with its strong mechanisms for compliance to accountability, transparency, and consent regarding data processing, storage, and transfer.

However, compliance to the GDPR regulations poses certain difficulties, including the very subject matter, the legal distinctions, and the constellation surrounding general legal language, data protection laws, which is intricate and in an ongoing process of development. Most legal texts written in privacy policies come with special terms and other complicated statements, which many people may not understand. Most compliance checks have been done manually and this requires a lot of time, and is prone to errors, thus being expensive for the organizations. However, the GDPR is not cast in stone, it is dynamic, meaning that it is subject to change through amendments or reinterpretation to mention but a few; as a result, organizations are forced to be on their guard all the time to avoid non-compliance [Sirur, Nurse, Webb, 2018].

To overcome these trends, Natural Language Processing (NLP) is a solution to help automate and interpret legal documents more accurately in compliance. NLP based techniques helps the organization in analyzing the policy clauses, extracting information from it and evaluating the GDPR compliance which helps in effective compliance, better understanding of the policy clauses and hence enhanced transparency and improved data protection. Likewise, when a set of regulatory changes the law it shouldn't be a problem for NLP models to respond to these changes and as such, organizations can adjust to the new legal requirements with ease [Aberkane, Poels, Broucke, 2021; Cejas et al., 2023]. The GDPR compliance focus is the article 5 of the GDPR that highlights six principles of processing personal data and one responsibility of the controller.

This preliminary study explores how NLP can assist in GDPR compliance, examining three core research questions:

- How effective are NLP models in automating GDPR compliance checks within organizational data privacy policies?
- What limitations do current NLP technologies face in interpreting and enforcing GDPR requirements, and how can these be addressed?
- What role can NLP-powered tools play in supporting compliance officers and legal experts in maintaining GDPR adherence?

Our objectives are to capture the degree that NLP can support with automating GDPR compliance and perform a preliminary study. This preliminary work gives a research inspiration but merely scratches the surface. As a preliminary study, this paper establishes some foundational insights into model strengths and limitations. Further research will be needed to optimize these methods or prove the theories on multiple data sets, and apply them to practical usage.

Methods

The essence of the methodology is the training of the multi-label classification on the GDPR principles, the generation of compliance reports, the usage of the metrics which are oriented specifically on the evaluation of the model's performance and its precision.

In this study, two different data sets were used: OPP-115 and ACL Coling. They were utilized to train and verify the performance of a couple of NLP models, including SBERT, BERT, and GPT2, which are one of the most powerful languages processing technologies.

Dataset acquisition

Two datasets were retrieved from Usable Privacy Policy [Usable Privacy Policy Project]. The combination of the two datasets allows for more reliable results from training and testing the models.

OPP-115 Dataset

OPP-115 which comes from Online Privacy Policy Project is a dataset of 115 policy policies gathered in 2016 from various websites. It provided a good starting point for the training on NLP models on GDPR compliance. The policies in this dataset are all annotated in detail by a set of defined categories that do align with the principles provided under the GDPR article 5.

A sample 0.1 of how an annotation looks like can be seen below. The text highlighted in blue shows the data type under the category and the extracted sentence that pertains to the category, respectively.

Sample 0.1: Snippet of annotated data from OPP-115 dataset

Third Party Sharing/Collection

{ “Third Party Entity”: {“endIndexInSegment”: 614, “startIndexInSegment”: 76, “selectedText”: “We may disclose or share individual, nonpersonally identifiable information and aggregate information in any manner other than that described herein that we deem appropriate or necessary, “value”: “Unnamed third party”}, . . . }

http://www.theatlantic.com/privacy-policy/

ACL Coling Dataset

ACL Coling Dataset was used mainly for the evaluation purposes. It has a different range of 1,010 legal documents collected in 2014 in xml format. The corpus itself was made for the Computational Linguistics Conference and a sample of how the data looks like is shown below:

Sample 0.2: Sample of data structure from ACL-Coling dataset

<POLICY> modification_date=“ ”
policy_url=“http://earthclinic.com/privacy_policy.html”
website_category=“Health” website_index=“098”
website_url=“earthclinic.com”
<SECTION>
<SUBTITLE />
<SUBTEXT>
Earth Clinic, LLC Privacy Policy Agreement The privacy of our visitors to the Earth Clinic web site, www.earthclinic.com (“Earth Clinic, LLC”) is important to us. Because we gather certain types of information about our users, we want to help you understand the terms and conditions surrounding the collection and use of that information. This privacy statement discloses the types of information we gather,

how we use it, and how to correct or change it. These privacy practices apply to the Web site that you were viewing when you clicked through to this policy, which is operated directly by us. Here is information on what types of personal information we receive and collect when you use visit the Earth Clinic LLC web sites and how we safeguard your information. We never sell your personal information to third parties. . .
</SUBTEXT>
</SECTION>
<SECTION>
<SUBTITLE> Privacy Certifications </SUBTITLE>
<SUBTEXT> </SUBTEXT>
</SECTION>
</POLICY>

From the sample 0.2, the text highlighted in pink is the root of the given policy, inside it are sections highlighted in yellow. Each section has a subtitle highlighted in green, and a subtext highlighted in purple.

Attribute	OPP-115	ACL Coling
Focus	Privacy Policy Analysis	Computational Linguistics Research
Content Type	Website Privacy Policies	Website Privacy Policies
Annotations	Data Practices (e. g., collection, sharing)	None
Use Case	Training models to analyze privacy policies	Broad NLP and linguistic research
Data Format	Annotated Text	Plain Text, PDF
Accessibility	Restricted Access	Open Access
Size	115 policies	1,010
Year of Release	2016	2013 & 2014
Maintained by	Usable Privacy Policy Project	Association for Computational Linguistics

Data preprocessing

For any text analysis how effective the data preprocessing goes a long way which affects the performance of the models. Especially in legal texts, it is very important as text representation accuracy plays a significant role. This section discusses the preprepocessing techniques implemented on both the OPP-115 and ACL Coling datasets to make them suitable for continuing the preliminary studies.

The OPP-115 dataset required extensive text cleaning especially because to get the policy with its equivalent label, we needed to go through the annotated version and put sentences together to get the entire privacy policy. The preprocessing steps included text cleaning, tokenization and lemmatization.

The ACL Coling dataset requires a slightly different approach to preprocessing. Given the nature of XML structure, the process begins with parsing the XML documents to extract properly the textual content. So a code was written to identify and read the specific elements in the XML hierarchy. These are the section subtexts, because they are where the relevant legal text is stored.

Once the text is extracted, it undergoes some further cleaning steps: normalization, whitespace removal, tokenization, lemmatization and punctuation removal. These preprocessing steps are done in order to maximize how efficient the NLP models used are in this preliminary study. And that is achieved by providing clean, consistent, and meaningful text data.

Mapping GDPR principles

The mapping of GDPR principles to the categories of the OPP-115 dataset is the next step in the methodology. Because we want to classify policies as compliant or not based on if they follow the

GDPR principles. This section outlines how and why each category from the OPP-115 dataset map to corresponding GDPR principles from Article 5.

The OPP-115 dataset categorizes listed in the data acquisition subsection map to the GDPR principles as demonstrated by [Poplavska et al., 2020]. The mapping is shown in Figure 1.

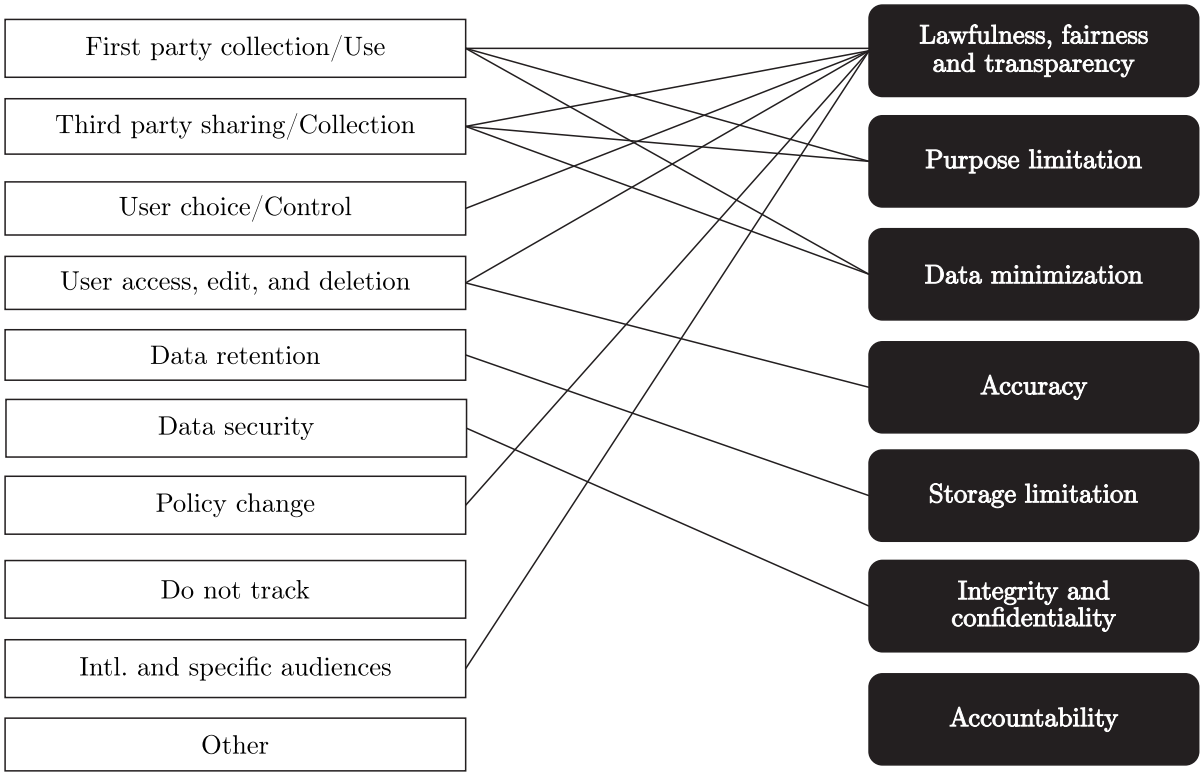

Figure 1. Mapping of OPP-115 categories to GDPR article 5 principles [Poplavska et al., 2020]

This mapping process not only guides the training of NLP models but also helps in structuring the compliance checks.

Multi-label classification

For this research, the multi-label classification method was employed to define the measure of noncompliance of each chunk of text, single and numerous SENTs, or the entire POLICY with the seven principles outlined under Article 5 of the GDPR.

Nevertheless, within the framework of GDPR, every single statement or even the entire policy is assessed not only in terms of compliance or noncompliance but is considered as to how many principles of the GDPR are simultaneously violated by it. This is a real-world situation of a single privacy policy seeming to address some of the GDPR principles and disregard the others. Hence, each text unit is connected with a numeric vector of length seven, which captures the admittance to the seven principles of the GDPR. Every component of the vector is binary, equal to either 0 or 1; small number 1 means compliance with the specific GDPR principle, while number 0 refers to violations.

Labels

The labels for this classification task are derived from the GDPR principles [Nicolaidou, Georgiades, 2017], which are: Lawfulness, Fairness, and Transparency, Purpose Limitation, Data Minimization, Accuracy, Storage Limitation, Integrity and Confidentiality and Accountability. Each of these principles forms a dimension in the label vector for each data point. A sentence or policy is

annotated with “1” for a principle if it aligns with the GDPR requirements for that principle, and “0” otherwise.

Compliance reporting

Effective GDPR compliance requires not only identifying potential noncompliance but also clearly reporting these findings. This section outlines the format used in this research to transform model outputs into comprehensive compliance reports.

Report format

The final compliance report is structured to provide clear and actionable insights. Each report has the following structure:

Sample 0.3: Example of a compliance report

GDPR Compliance Report for Policy XYZ
Summary: Noncompliant with 2 out of 7 principles evaluated.

Detailed Findings:
1. Data Minimisation: Compliant.
2. Integrity and Confidentiality: Noncompliant.
Example: “User data may be stored indefinitely for analytics.”

Recommendations:
– Review the data retention policy to align with the “storage limitation” principle.

Thus, the presented compliance reporting framework guarantees that the results of the models are not only understandable but also practical, which makes them significant for organizations that need proper GDPR compliance.

Implementation

Each model was trained on both the sentence level and entire policy level using the labeled OPP-115 dataset. The models were then saved and subsequently tested on unlabeled policies from the ACL Coling dataset to assess their generalizability and performance.

Data preparation

Initially, sentences and the whole policies themselves were annotated with binary vectors which can describe the compliance concerning seven GDPR principles, and make it possible to train separate models for the practical analysis at the sentence- and policy-levels. Each of the data was tokenized and encoded appropriate to the needs of the given model; for example, SBERT, BERT, and GPT2 work under conditions that require data to be formatted in a certain way.

Label assignment and distribution

Notably, each sentence or policy was described with a binary 7-vector where each component is 1 at the position relating to the corresponding GDPR principle. Each of them is brought to the vector; if the value of the element is 1, then the organization adheres to the corresponding principle, and if it is 0, then the organization does not adhere to this principle. The distribution of these labels across the dataset is shown in Table 1.

Before training, the OPP-115 dataset was prepared at two different granularities:

Table 1. Distribution of compliance with GDPR Principles

GDPR principle	Sentence instances	Policy instances
Lawfulness, Fairness, and Transparency	8460 compliant	115 compliant
Purpose limitation	6209 compliant	115 compliant
Data minimization	6209 compliant	115 compliant
Accuracy	646 compliant	90 compliant
Storage limitation	396 compliant	76 compliant
Integrity and confidentiality	1000 compliant	102 compliant
Accountability	0 compliant	0 compliant

1. Sentence level: In the case of the privacy policies, each sentence was annotated with regard to how well it adhered to the presented GDPR principles, which comprised the training data for the sentence-level models.
2. Entire policy level: Full privacy policies were categorized based on the GDPR compliance, even or odd, as a whole to train the policy-level models.

Model training and saving

As part of model training, it was necessary to set up structures for neural networks since the documents' embeddings are preprocessed using SBERT, BERT, and GPT2 models. All of the above models were trained using OPP-115 dataset split in training and validation sets. After training all the models were serialized in a standard way.

Related works

The topic of compliance in the past few years, and more specifically in relation to the GDPR as an example of data protection legislation, one cannot fail to notice the seemingly "revolutionary" changes that have been brought about by the incorporation of NLP capabilities [Aberkane, Poels, Broucke, 2021; Hamdani et al., 2021; Nazarenko, Lévy, Wyner, 2021]. As a result of this evolution, NLP has emerged into the upcoming area for automating diverse processes in accomplishing the data checking and privacy policies based on the GDPR and other data protection laws [Bonatti et al., 2020; Galle, Christofi, Elsahar; Amaral et al., 2021].

Combining NLP possibilities with legal compliance activities is a revolution in comprehending and dealing with regulatory requirements and legal documents. Typically, data protection compliance has been a time-consuming and expensive affair that has entailed paper-based sifting and analysis of vast legal texts [Li et al., 2020; Mori et al., 2022; Qamar, Javed, Beg, 2021]. Though there is not much information about its use in handling compliance, the use of NLP has brought efficiency and scalability in the organization's use of Machine learning and natural language understanding to handle compliance work [Sousa, Kern, 2023; Srinath, Wilson, Giles, 2020; Silva et al., 2020a].

Overview

Thus, this extensive literature review aims at discussing key publications in this field and highlights various methodologies, models, and approaches that rely on NLP to analyze the complexity of GDPR compliance and other legal documents. Looking at the specifics of data protection and privacy legislations [Alattas et al., 2022; Voigt, von dem Bussche, 2017], data protection officers and academic pioneers have developed brand new approaches that concern the difficulties in corresponding compliance procedures and the clarification of data practices in policy documents such as privacy policies, data processing agreements, and regulation requirements [Harkous et al., 2018; Leone, Di Caro, 2020; Müller et al., 2019].

Key areas of focus within this literature review include:

- Compliance checking: There are well and advanced models and frameworks that describe how compliance checking of the GDPR requirements can be automated. These models act as legal advisors that employ NLP strategies to analyze legal texts and determine whether an organization conforms to the requirements of legal precedents [Alattas et al., 2022; Amaral Cejas, Abualhaija, Briand, 2023; Torre et al., 2020].

- Privacy policy analysis: Corpora as well as tools like PrivaSeer have emerged as powerful NLP solutions of analyzing policies and have greatly helped in large scale collection of data extraction as well as classification [Arora et al., 2022; Bokaie Hosseini et al., 2020; Harkous et al., 2018]. These performances may be pursued to increase or improve the current state of transparency together with improving the ability of the users in making proper choices regarding their right to privacy and protection of their data.

- Semantic annotation: Previous attempts at indexing and annotating legal texts with semantics have made it possible to achieve enhanced perform for search as effectively as for details mining [Ling et al., 2023; Sánchez, Viejo, Batet, 2021; Silva et al., 2020b]. Through adding metadata and semantic tags into legal texts, the scholars have created opportunities for practically working in the future to improve the availability and understanding of legal regulations.

Thus, it is within this framework of utilizing these pioneers in the field of NLP that this literature review aims at uncovering the possibility of the NLP methodology in the GDPR context and data protection. Thus, by integrating various methodologies and approaches, the scholars try to enhance innovation, increase the roles of methods' transparency, and help organizations to manage the challenges of the constantly changing legal environment effectively [Li et al., 2020; Leone, Di Caro, 2020; Amaral et al., 2022].

Challenges in leveraging NLP for GDPR compliance

Nevertheless, there are some issues that should be noted, proving that further research and development in the field of NLP for GDPR, as well as legal text analysis in general, should continue [Alattas et al., 2022; Voigt, von dem Bussche, 2017]. These issues must be further solved for the purpose of providing dedicated solutions that are reliable and can easily address various aspects related to regulation and legal documentation [Aberkane, Poels, Broucke, 2021; Amaral et al., 2021; Del Alamo et al., 2022].

1. Dataset limitations

One of the issues is the fact that these models need to be currently trained on larger and more diverse datasets for best results — hence accessibility. Currently produced datasets may not be sufficient in providing the needed scope and depth in coverage to reflect the variations of the legal language use and the associated compliance situations. There is, therefore, a dire need to assemble relevant collections of overall legal works, legal codes, and linguistic differences [Li et al., 2020; Leone, Di Caro, 2020; Hamdani et al., 2021].

2. Adaptability to evolving legal frameworks

The legal frameworks in the case of NLP-based compliance solutions are GDPR and other data protection laws which emerge as the biggest problem due to how volatile they are. Legal texts change often, get interpolated, amended, or simply reinterpreted quite regularly, which requires models to be at par with these changes as and when they happen. NLP models that are flexible and dynamic enough to follow any changes to the legal specifications for compliance are needed [Poplavska et al., 2020; Mousavi Nejad et al., 2020; Sánchez, Viejo, Batet, 2021].

3. Exploration of newer NLP architectures

Although the current NLP architectures have shown a good level of achievement in multiple utilities such as compliance checks and legal textual analysis, researchers are still trying to discover new architectures or pre-trained models. The Generative models including GPT-3 and T5 can be used to produce natural responses and also for context-based legal documents comprehension. It would be interesting to try out these neural architectures and incorporate these architectures into tasks that deal with compliance; there could be much more potential for enhanced accuracy and efficiency here [Li et al., 2020; Liu et al., 2021; Giner-Miguelez, Gómez, Cabot, 2023].

Gap analysis

On the basis of these research gaps and the limitations mentioned in the present work to further develop the NLP for GDPR compliance field, the following strategies can be outlined for future research.

1. Experimentation with generative models

The method of using generative models should be extended in future studies with up-to-date generative models such as GPT-3 and T5. These models have enhanced abilities in NLP processing and generation that are especially appropriate for complex applications such as compliance, legal document review, and similar processes [Silva et al., 2020b; Sousa, Kern, 2023; Srinath, Wilson, Giles, 2020].

2. Extensive evaluations

It is critical to conduct more comprehensive investigations with respect to NLP-based compliance solutions and applications as to how well they work, how resilient they are, and how easily they can be scaled up. The evaluation framework should contain multiple datasets, metrics, and applications to give recommendations for utilizing the approach and information about its effectiveness [Bonatti et al., 2020; Amaral Cejas, Abualhaija, Briand, 2023; Del Alamo et al., 2022].

3. Interdisciplinary approaches

Another possible method for further research involves the focus on both legal knowledge incorporated into the engineering of compliance solutions and NLP methods to improve such solutions. Multidisciplinary collaborations with law enforcement, IT, compliance, and NLP specialists can work towards improving the current solutions by creating tools that are more aware of the specific issues within the complex sphere of compliance [Sleimi et al., 2018; Rahat, Long, Tian, 2022; Harkous et al., 2018].

In summary, the literature reviewed supports the centrality of NLP for the complex problems concerning GDPR regulation, privacy policy comprehension, and legal documents deciphering. Further research and experimentation have to be conducted to fill this gap and enhance the state of the art in the applied approaches for automated compliance checking and legal text analysis.

Discussion

This section delves into the performance evaluation of the machine learning models used in this research, namely, SBERT, BERT, and GPT2. The focus is on finding if they are efficient in GDPR compliance classification at both the sentence and policy levels.

Results

Overview of model performance

The models were evaluated based on accuracy, precision, recall, and F1-score; for example, as seen below, the charts for SBERT performance on policy level.

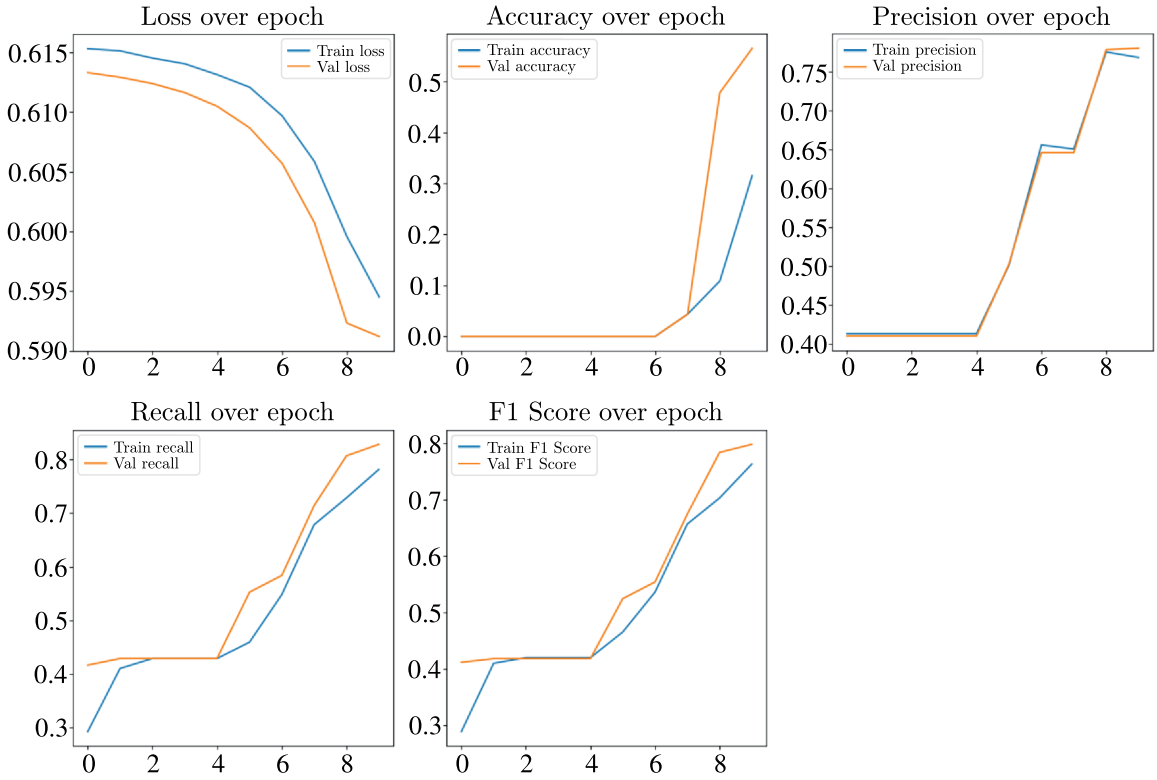

Figure 2. Metrics for SBERT performance on policy level

For the results, after looking at the explanation above we can see that SBERT performed the best across all the models practically even though in terms of metrics it was far from the best. Let's take a dive into how it works and predicts compliance for different sentences and privacy policies. We will start with checking compliance of sentences against a single principle from the GDPR article five: "Storage limitation".

Four sample sentences, two of which adhere to Storage Limitation and two of which don't were evaluated on the three models. Then moving on to test these models in a more practically evaluation based way, they were fed the sentences shown below. It was interesting to note that for single principle training, they all produced equivalent and correct results.

After seeing the result, I went ahead to confirm ChatGPT API and a sample response for a noncompliant and compliant sentence, respectively, were as follows:

From the sentence: "we collect two kinds of information related to you: (a) personally identifiable information ('pii'); and (b) nonpersonally identifiable information ('non-pii')." Does it comply with any of the principles from the GDPR article 5? If it does, then which is the main one?

Hello! Yes, the sentence you provided does comply with principles from Article 5 of the General Data Protection Regulation (GDPR). Specifically, it aligns most closely with the principle of "Purpose limitation" and "Data minimization"...

Table 2. Compliance analysis for privacy principles using NLP models

Principle	Sentence	ChatGPT API	SBERT	BERT	GPT2
Storage limitation	We collect two kinds of information related to you: (a) personally identifiable information (“PII”); and (b) nonpersonally identifiable information (“non-PII”).	Purpose limitation	False	False	False
	Furthermore, some information is never completely removed from our databases due to technical (e. g., information that is stored in our “back-up” systems) and certain legal constraints.	Storage limitation	True	True	True
	This policy states that user data will be stored for a maximum of two years.	Storage limitation	True	True	True
	We reveal only the last four digits of your credit card numbers when confirming an order.	Data minimization	False	False	False

Meanwhile for the second sentence that was compliant, the response from ChatGPT was:

> . . . The sentence appears to relate most directly to the **Storage limitation** principle. This principle requires hat personal data be kept no longer than necessary for the purposes for which it is processed. However, the sentence explains why some data cannot be completely removed due to technical and legal constraints, which acknowledges that there are exceptions to the complete erasure of personal data. . .

After this little success, it was relevant to move on and try it out in a more practical setting. There was a privacy policy that was fed to the model and it had to extract the sentences that made it classify this policy as adhering to **Storage limitation**.

Specifically, the highlighted sections show the adhering sentences as printed by the model and they comply with the definition given by the GDPR of this principle.

So far this has been just training and testing on one principle. Following this, let’s look at the results of the multilabel classification for all 7 principles on the sentence level. In this case all three of the models used a relatively high threshold of 0.8 and had the metrics as seen in Table 3.

Table 3. Comparison of metrics for multi-principle classification at the sentence level

Metric	SBERT	BERT	GPT2
Accuracy	0.587	0.63	0.62
Precision	0.56	0.70	0.72
Recall	0.43	0.50	0.48
F1 Score	0.44	0.55	0.54

Subsequently, these models were saved and tested on entire policies. However, because they were trained on a sentence level granularity, the privacy policies were split into sentences before being analyzed by the models and finally a compliance report was generated highlighting the principles covered with sample sentences and which principles need to be worked on.

Next for the policy level, SBERT only provided results when the threshold was as low as 0.5. The metrics for all models can be seen in Table 4.

Figure 3. SBERT Compliance Report for Sentence Level

Table 4. Comparison of metrics for multi-principle classification at the policy level

Metric	SBERT	BERT	GPT2
Accuracy	0.57	0.26	0.48
Precision	0.78	0.77	0.82
Recall	0.83	0.75	0.75
F1 Score	0.80	0.73	0.78

At the policy level, the reports were not created because the models were not able to extract sentences and could only classify the entire policies as trained to do. The results were instead in the following format and just like with the sentence level granularity and the single policy classification, SBERT provides more detailed results.

All in all, there are certainly areas in policy matching that SBERT will perform well owing to the superior quality of the sentence embeddings to capture semantic similarities across different texts compared to token based models. This allows SBERT to retain context over multiple sentences which is important when gauging larger compliance in GDPR policies. While there is lower document level coherence in cases of BERT and better text generating capabilities in GPT2, SBERT is assumed to understand the sentences well particularly and to have added advantages of merging legal language and accurately identify compliance, thus showing better recall and F1-scores.

Challenges encountered

The following are some of the challenges which the models experienced while in the process of the research along with the kind of approaches that were taken to handle the said challenges.

Class imbalance

One major issue that was experienced was the issue of class imbalance, especially regarding GDPR Principle 7 (Accountability) where there were no passing cases. This affected the training process because stratification occurred in a way that favored a certain class and so strategies such as class weights had to be applied to address this issue. The weighting of the trainers mitigated the impact of one of the classes during the training of the model.

Dataset preparation and preprocessing

In the aspect of preparing and preprocessing the dataset, there were multiple challenges:

- Annotations to labels conversion: The pre-existing data set had annotations instead of labels in it. Furthermore, such annotations were aligned with categories of OPP-115 instead of GDPR principles.
- Manual mapping: Finally, a paper that contained the mapping between OPP- 115 categories and GDPR principles was used to solve this problem as well.
- Sentence formation: Threshold to construct sentences was somewhat difficult to construct with regards to the number of words that constitutes as a sentence and handling the issue of duplicates.

Threshold selection and hyperparameter tuning

Selecting the right thresholds was also an issue. Some of the models had the maximum probability of assigning to a prediction as 0.53 does not reflect the criteria of 0.8 which was wanted for the final step of prediction. Hyperparameter tuning was used to optimize the model and search out the best parameters of the model such as learning rate, threshold and training epochs.

Model selection

Deciding between base models and legal-special models such as nlpaueb/legal-bert-base-uncased was another one of the problems. Some of the time, legal versions of models did not perform as well as hypothesized or result in any significant difference, and it was decided to stick with the base models for the tasks. The following table is to show the performance results of the legal versions of the models against the base models. It helps to compare which of the model versions is better suited for the given task.

Table 5. Comparison of Legal vs. Base Model for BERT

Metric	Base BERT		Legal BERT	
	Sentence level	Policy level	Sentence level	Policy level
Accuracy	0.62	0.56	0.43	0.13
Precision	0.72	0.72	0.73	0.70
Recall	0.45	0.85	0.34	0.65
F1 Score	0.52	0.77	0.44	0.67

Max length limitation of models

Another interesting problem faced was the restriction of the input of BERT and GPT2 models up to 512 tokens. This limitation meant that it was not possible to show the models full policies at once and therefore text always had to be split into sub-sections. While this approach proved beneficial in terms of the models' ability to analyze the data, it also had the side effect of limiting context and continuity between segments, which could decrease the validity and reliability of the outcomes.

Resource and financial constraints

Several resource and financial constraints were encountered:

- OpenAI API costs: Since the OpenAI embeddings API is not free, it incurred a lot of expenses in trying to make the deep learning model work. This financial constraint was one of the reasons why extensive use of this method could not be applied.
- Compute resources: Running models with GPU on Colab required payment for additional compute units. Moreover, due to long running times of the models getting results was a slow and costly process.

Conclusion

In this research, the focus was on using NLP to automate compliance checking for the GDPR. This undertaking is inspired by such factors as the growing difficulty in legal matters and the need to address data protection laws in organizations effectively. The comprehensive analysis involved leveraging state-of-the-art NLP models, including SBERT, BERT, and GPT2, across two granularity levels: on the one hand, they are at the sentence level and, on the other hand, at the entire policy level.

Summary of findings

Since the aim of this study was to evaluate the NLP models to determine the degree of compliance with GDPR in privacy policies, through rigorous experimentation and evaluation, several key findings emerged:

From the models evaluated, it can therefore be deduced that while all the models were fairly effective to a certain extent, SBERT fared best at the policy level by metrics, providing high accuracy and f1 score of compliance matters. Whereas on evaluation by practicality SBERT outperfromed all other levels across both granularities. There were also acceptable scores for BERT and GPT2 models; BERT showed better scores by metrics and they did major in analyzing relationships in the text. They tended to provide more tightly narrowed down text analysis in their tests on actual policies. The following scores summarize the performance where sentence level maximums are highlighted as blue and policy level maximums are highlighted as orange:

- SBERT:
  - Sentence level: Accuracy: 0.58, Precision: 0.56, Recall: 0.43, F1-score: 0.44;
  - Policy level: Accuracy: 0.57, Precision: 0.78, Recall: 0.83, F1-score: 0.80;
- BERT:
  - Sentence level: Accuracy: 0.63, Precision: 0.70, Recall: 0.50, F1-score: 0.55;
  - Policy level: Accuracy: 0.26, Precision: 0.77, Recall: 0.75, F1-score: 0.73;
- GPT2:
  - Sentence level: Accuracy: 0.62, Precision: 0.72, Recall: 0.48, F1-score: 0.54;
  - Policy level: Accuracy: 0.48, Precision: 0.82, Recall: 0.75, F1-score: 0.78.

Answering the research questions

This research set out to answer several key points, and the findings gotten give some answers and insights into the chosen research questions:

- How effective are NLP models in automating the identification of GDPR compliance issues within organizational data privacy policies? The findings show that NLP models are very efficient and accurate with primary attention to the SI model including SBERT and BERT for comprehending compliance.
- What are the limitations of current NLP technology in interpreting and enforcing GDPR compliance, and how can these limitations be addressed? The primary drawbacks consist of computational complexity and the necessity of utilizing less cognitively complex models, along with the issues of interpretability. Overcoming these limitations can be achieved by enhancing the computational speed, making the models more available and coming up with ways through which models can be easily understood.

- What role can NLP-powered tools play in supporting compliance officers and legal experts in maintaining GDPR compliance? It is also demonstrated that NLP-powered tools can assist compliance officers by automating those tasks as proper identification of compliance problems in organisation, which decreases the volume of workload of the officer and increases the level of reliability of compliance checks. These tools could be helpful in compliance which in its turn would free up the legal experts' time to perform more sophisticated tasks.

Future work and improvements

While this research has focused on automating GDPR compliance checks using several NLP models like SBERT, BERT, and GPT2, and it has revealed specific directions for future research and improvement. The possible future studies and improvements that can be made are indicated in this section.

Alternative solutions to challenges

- Class imbalance: Since the class imbalance results from the feature of GDPR article 5 where the seventh principle is only for the controller not the data, this issue is an exemption. However, trying other deep learning loss functions especially for imbalanced datasets such as focal loss that are trained to reduce bias of majority classes could solve class imbalance.
- Dataset preparation and preprocessinge: Dynamically defining the threshold based on the distibution of the dataset's sentence length could make clearer segmentation and avoid duplicate detection of similar sentences.
- Threshold selection and hyperparameter tuning: Implementing a more adaptive threshold depending on the model confidence could increase flexibility in predictions.
- Model selection: Developing hybrid models that utilize both the benefits of legal-specific models and base models could produce a model with the strengths of both model types.

Enhancing model performance

Future research could improve the performance of the NLP models and this could be achieved through several axes such as:

- Data augmentation: Making the dataset more diverse and recent, because the current dataset is from before 2020. In addition, the training data set can be augmented upon to create richer datasets.
- Fine-tuning: Increased fine-tuning of models on larger amounts of this data in order to better interpret legal language and text.
- Model architecture: Trying new architectures and putting together different models to get higher accuracy, precision, recall and the F1 score.

Real-time compliance monitoring

Another important area for the future work is the procedures to be used to apply real-time compliance monitoring systems effectively. Such systems could be able to constantly search for and assess new policies or modifications to existing ones to make sure of compliance to GDPR, progressively. This involves:

- Automated pipelines: Developing pipelines that can be integrated with actual industry data flows. This will allow the compliance checks to be smoother; this is checked as soon as policies are created or updated.

- Alert systems: Setting up alerts to inform the organization of the areas that might not be in compliance in real-time.

Broadening the scope of compliance checks

Extending the application of the compliance check itself to other regulations concerning data privacy like CCPA, HIPAA, and the other laws is another potential avenue for future research. This includes:

- Multi-regulation frameworks: Developing frameworks that can assess compliance with multiple regulations simultaneously.
- Cross-jurisdictional analysis: Facilitating easy comparison of compliance across jurisdictions so as to make compliance checking better for multinational corporations.

User-centric enhancements

Future developments should also aim at enhancing the way the tool can adapt to the needs of the compliance professionals, organizations or legal users. This involves:

- Customization: Allowing users to have tools that allow them to customize compliance checks based on their specific business needs.
- Integration with existing tools: Integration with other current legal and compliance tools that are already being used by the organizations.

All in all, despite the work of this research providing a good foundation for employing NLP models to automate GDPR checks, there are several possibilities for further research and manipulation. In these aforementioned areas, it is possible to enhance the potential of new tools and improve the experiences concerning the management of data protection regulations.

References

Aberkane A.-J., Poels G., Broucke S. V. Exploring automated GDPR-compliance in requirements engineering: a systematic mapping study // IEEE Access. — 2021. — Vol. 9. — P. 66542–66559. — DOI: 10.1109/ACCESS.2021.3076921

Alattas H. T., Almassary F. M., AlMahasheer N. R., Alammari R. M., Alswaidan H. A., Nagy N. M., Almoqbel M. A., Alharthi S. A. Extract compliance-related evidence using machine learning // 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN). — 2022. — P. 537–542. — DOI: 10.1109/CICN56167.2022.10008324

Amaral O., Abualhaija S., Sabetzadeh M., Briand L. A model-based conceptualization of requirements for compliance checking of data processing against GDPR // 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW). — 2021. — P. 16–20. — https://ieeexplore.ieee.org/document/9582337

Amaral O., Abualhaija S., Torre D., Sabetzadeh M., Briand L. C. AI-enabled automation for completeness checking of privacy policies // IEEE Transactions on Software Engineering. — 2022. — Vol. 48, No. 11. — P. 4647–4674. — DOI: 10.1109/TSE.2021.3124332

Amaral Cejas O., Abualhaija S., Briand L. ML-based compliance verification of data processing agreements against GDPR // 2023. — https://orbilu.uni.lu/handle/10993/55408

Arora S., Hosseini H., Utz C., Bannihatti Kumar V., Dhellemmes T., Ravichander A., Story P., Mangat J., Chen R., Degeling M., Norton T., Hupperich T., Wilson S., Sadeh N. A tale of two regulatory regimes: creation and analysis of a bilingual privacy policy corpus // Proceedings of the Thirteenth Language Resources and Evaluation Conference. — 2022. — P. 5460–5472. — https://aclanthology.org/2022.lrec-1.585

Bokaie Hosseini M., Pragyan K. C., Reyes I., Egelman S. Identifying and classifying third-party entities in natural language privacy policies // Proceedings of the Second Workshop on Privacy in NLP. — 2020. — P. 18–27. — https://aclanthology.org/2020.privatenlp-1.3

Bonatti P. A., Kirrane S., Petrova I. M., Sauro L. Machine understandable policies and GDPR compliance checking // KI – Künstliche Intelligenz. — 2020. — Vol. 34, No. 3. — P. 303–315. — DOI: 10.1007/s13218-020-00677-4

Cejas O. A., Azeem M. I., Abualhaija S., Briand L. C. NLP-Based automated compliance checking of data processing agreements against GDPR // IEEE Transactions on Software Engineering. — 2023. — Vol. 49, No. 9. — P. 4282–4303. — DOI: 10.1109/TSE.2023.3288901

Del Alamo J. M., Guaman D. S., García B., Diez A. A systematic mapping study on automated analysis of privacy policies // Computing. — 2022. — Vol. 104, No. 9. — P. 2053–2076. — DOI: 10.1007/s00607-022-01076-3

Galle M., Christofi A., Elsahar H. The Case for a GDPR-specific annotated dataset of privacy policies.

Giner-Miguelez J., Gómez A., Cabot J. DataDoc analyzer: a tool for analyzing the documentation of scientific datasets // Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. — 2023. — P. 5046–5050. https://dl.acm.org/doi/10.1145/3583780.3614737

Hamdani R. E., Mustapha M., Amariles D. R., Troussel A., Meeùs S., Krasnashchok K. A combined rule-based and machine learning approach for automated GDPR compliance checking // Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. — 2021. — P. 40–49. — https://dl.acm.org/doi/10.1145/3462757.3466081

Harkous H., Fawaz K., Lebret R., Schaub F., Shin K. G., Aberer K. Polisis: automated analysis and presentation of privacy policies using deep learning // 2018. — P. 531–548. — https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

Leone V., Di Caro L. The role of vocabulary mediation to discover and represent relevant information in privacy policies // Frontiers in artificial intelligence and applications / Eds.: S. Villata, J. Harašta, P. Křemen. — IOS Press, 2020. — http://ebooks.iospress.nl/doi/10.3233/FAIA200851

Li Z. S., Werner C. M., Ernst N. A., Damian D. GDPR compliance in the context of continuous integration // ArXiv. — 2020. — https://www.semanticscholar.org/paper/71e16573d39360b98306b3bfa5482c10b4e73746

Ling Y., Wang K., Bai G., Wang H., Dong J. S. Are they toeing the line? Diagnosing privacy compliance violations among browser extensions // Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering. — 2023. — P. 1–12. — https://dl.acm.org/doi/10.1145/3551349.3560436

Liu S., Zhao B., Guo R., Meng G., Zhang F., Zhang M. Have you been properly notified? Automatic compliance analysis of privacy policy text with GDPR article 13 // Proceedings of the Web Conference 2021. — 2021. — P. 2154–2164. — https://doi.org/10.1145/3442381.3450022

Mori K., Nagai T., Takata Y., Kamizono M. Analysis of privacy compliance by classifying multiple policies on the web // 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC). — 2022. — P. 1734–1741. — https://ieeexplore.ieee.org/document/9842614/

Mousavi Nejad N., Jabat P., Nedelchev R., Scerri S., Graux D. Establishing a strong baseline for privacy policy classification // ICT systems security and privacy protection / Eds.: M. Hölbl, K. Rannenberg, T. Welzer. — Cham: Springer International Publishing, 2020. — P. 370–383. — DOI: 10.1007/978-3-030-58201-2_25

Müller N. M., Kowatsch D., Debus P., Mirdita D., Böttinger K. Text, speech, and dialogue / Ed. K. Ekštein. — Cham: Springer International Publishing, 2019. — P. 151–159. — DOI: 10.1007/978-3-030-27947-9_13

Nazarenko A., Lévy F., Wyner A. A pragmatic approach to semantic annotation for search of legal texts — an experiment on GDPR // Frontiers in artificial intelligence and applications / Ed.: E. Schweighofer. — IOS Press, 2021. — https://ebooks.iospress.nl/doi/10.3233/FAIA210313

Nicolaidou I. L., Georgiades C. The GDPR: new horizons // EU internet law: regulation and enforcement / Eds.: T.-E. Synodinou, P. Jougleux, C. Markou, T. Prastitou. — Cham: Springer International Publishing, 2017. — P. 3–18.

Poplavska E., Norton T. B., Wilson S., Sadeh N. M. From prescription to description: mapping the GDPR to a privacy policy corpus annotation scheme // International Conference on Legal Knowledge and Information Systems. — 2020. — https://api.semanticscholar.org/CorpusID:229377855

Qamar A., Javed T., Beg M. Detecting compliance of privacy policies with data protection laws. — 2021.

Rahat T. A., Long M., Tian Y. Is your policy compliant? A deep learning-based empirical study of privacy policies' compliance with GDPR // Proceedings of the 21st Workshop on Privacy in the Electronic Society. — 2022. — P. 89–102. — https://dl.acm.org/doi/10.1145/3559613.3563195

Sánchez D., Viejo A., Batet M. Automatic assessment of privacy policies under the GDPR // Applied Sciences. — 2021. — Vol. 11, No. 4. — P. 1762. — DOI: 10.3390/app11041762

Silva P., Gonçalves C., Godinho C., Antunes N., Curado M. Using natural language processing to detect privacy violations in online contracts // Proceedings of the 35th Annual ACM Symposium on Applied Computing. — 2020a. — P. 1305–1307. — https://doi.org/10.1145/3341105.3375774

Silva P., Gonçalves C., Godinho C., Antunes N., Curado M. Using NLP and machine learning to detect data privacy violations // IEEE INFOCOM 2020 – IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). — 2020b. — P. 972–977. — https://ieeexplore.ieee.org/abstract/document/9162683

Sirur S., Nurse J. R. C., Webb H. Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR) // Proceedings of the 2nd International Workshop on Multimedia Privacy and Security. — 2018. — P. 88–95. — https://doi.org/10.1145/3267357.3267368

Sleimi A., Sannier N., Sabetzadeh M., Briand L., Dann J. Automated extraction of semantic legal metadata using natural language processing // 2018 IEEE 26th International Requirements Engineering Conference (RE). — 2018. — P. 124–135. — https://ieeexplore.ieee.org/document/8491129?denied=

Sousa S., Kern R. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing // Artificial Intelligence Review. — 2023. — Vol. 56, No. 2. — P. 1427–1492. — DOI: 10.1007/s10462-022-10204-6

Srinath M., Wilson S., Giles C. L. Privacy at scale: introducing the PrivaSeer corpus of web privacy policies // arXiv. — 2020. — http://arxiv.org/abs/2004.11131

Torre D., Abualhaija S., Sabetzadeh M., Briand L., Baetens K., Goes P., Forastier S. An AI-assisted approach for checking the completeness of privacy policies against GDPR // 2020 IEEE 28th International Requirements Engineering Conference (RE). — 2020. — P. 136–146. — https://ieeexplore.ieee.org/abstract/document/9218152

Usable Privacy Policy Project // https://www.usableprivacy.org/data

Voigt P., von dem Bussche A. The EU general data protection regulation (GDPR). — Berlin, Heidelberg: Springer, 2017.