Leveraging Pre-trained LLMs for GDPR Compliance in Online Privacy Policies

Giovanni Ciaramella[1,2,*], Luca Petrillo[1,2,*], Margaret Varilek[1,*], Francesco Mercaldo[5,2], Giovanni Comandé[4] and Fabio Martinelli[3]

[1]*IMT School for Advanced Studies Lucca, Lucca, Italy*
[2]*Institute for Informatics and Telematics of CNR, Pisa, Italy*
[3]*Institute for High Performance Computing and Networking of CNR, Rende, Italy*
[4]*Sant'Anna School of Advanced Studies, Pisa, Italy*
[5]*University of Molise, Campobasso, Italy*

Abstract

This article explores the use of Large Language Models (LLMs) to determine if online privacy policies comply with the General Data Protection Regulation (GDPR) since privacy policies do not always adhere to all relevant GDPR requirements. This paper proposes a method to classify privacy policies as compliant or not with a single duty within Article 13(2)(b) of the GDPR, which mandates that data subjects be informed of their right to rectification or erasure of personal data. To address that, we employed several LLMs such as BERT-base-uncased, roBERTa-base, distilBERT-base-uncased, t5-base, and ERNIE-2.0-base-en on a dataset built by the authors from European websites domains. Moreover, once the dataset was built, a legal expert from our research team manually classified a set of privacy policies to perform the contextual sentence similarity task. As the final step, we employed a set of unseen privacy policies to test models, obtaining interesting results demonstrating moderate accuracy in identifying compliant phrases using these thresholds. Future research could include expanding the analysis to encompass other GDPR requirements and refining the models.

Keywords

LLM, Privacy Policy, Compliance, Legal, AI, Cybersecurity

1. Introduction

In the European Union, privacy policies communicated by the data controller or processor to inform data subjects visiting a website cannot easily be evaluated in terms of "compliant" versus "non-compliant" by a human being. There may be aspects within a policy that might be compliant, and aspects which are likely to be considered not aligned with the General Data Protection Regulation (EU) 2016/679. The nature of GDPR compliance or privacy policy compliance are not black and white concepts to be determined about the entirety of the policy based on the face of the text of a policy alone. Therefore, to be "checkable" this bigger goal had to be broken down into smaller, verifiable steps. GDPR article 13(2)(b) (duty to inform the data subject of the right to rectification or erasure) was chosen to be tested as a first experiment. Since this task can be challenging for a human being, it can also be difficult for an AI model. However, it is still possible to leverage powerful models like the Large Language Model (LLMs) to assist humans in performing this time-consuming task. These models can be trained to dissect the text of privacy policies into specific terms and clauses and flag those which are relevant to selected articles of the GDPR. Additionally, they can analyze the phrasing within privacy policies and compare this to the language of the GDPR. They can spot inconsistencies, missing information or overly vague language that could fall short of compliance benchmarks. LLMs have the advantage of working with many texts in a short amount of time, making it possible to look at several privacy policies

Joint National Conference on Cybersecurity (ITASEC & SERICS 2025), February 03-8, 2025, Bologna, IT
*Corresponding author.
giovanni.ciaramella@imtlucca.it,giovanni.ciaramella@iit.cnr.it (G. Ciaramella); luca.petrillo@imtlucca.it,luca.petrillo@iit.cnr.it (L. Petrillo); margaret.varilek@imtlucca.it (M. Varilek); francesco.mercaldo@unimol.it (F. Mercaldo); giovanni.comande@santannapisa.it (G. Comandé); fabio.martinelli@icar.cnr.it (F. Martinelli)

from various organizations or services at once. This scalability is helpful for big companies looking to ensure compliance over separate platforms, such as Google, which operates various services in Europe: Google Search, YouTube, and Google Drive, each requiring adherence to different privacy policies. Using these capabilities, the goal is that eventually enterprises could utilize LLMs to support their compliance activities, mitigate risks, and confirm that their privacy agreements meet the jurisdiction's legal requirements, *e.g.*, GDPR. We choose most suitable LLM to identify the language corresponding to the roBERTa-selected text in privacy policies. This would then allow testing whether a sentence appears "compliant" with a single requirement. In order to achieve this goal, we employed state-of-the-art text embedding models (like all-mpnet-base-v2 [1] and all-distilroBERTa-v1 [2]). They excel at tasks that require understanding the semantic meaning of sentences, such as semantic similarity, clustering, and information retrieval. In addition, we relied on LLMs (like state-of-the-art BERT and variants) since they are designed for a broader range of tasks related to natural language understanding and generation, such as text generation, summarization, translation, and question answering. Preliminary results show that text classifiers returned "compliant" phrases, as verified by a human check. This step was performed by calculating three similarity scores (cosine similarity, dot product, and euclidean distance) between the manually selected privacy policy sentences and the Art. 13(2)(b). These metrics values are used to determine if a given sentence policy complies with the GDPR Article 13(2)(b). Future research could expand to additional GDPR requirements that make up a typical privacy policy, such as GDPR Art 13 (2)(d) which requires data subjects to be informed of the right to lodge a compliant with a supervisory authority. An important limit to bear in mind in this research process of using LLMs to check compliance of policies is that the quality of the result must be demonstrable in the words of the text alone.
The paper proceeds as follows: the next section shows an overview related to the state-of-the-art related to the adoption of AI techniques in the legal field; in Section 3, the proposed method is presented, while the results of the experimental evaluation performed are shown in Section 4 and, finally, conclusion and future research lines are drawn in the last section.

2. Related Work

Over the years, Artificial Intelligence has played a crucial role in several fields. AI profoundly impacts, primarily through Natural Language Processing (NLP) [3, 4] and Large Language Models (LLMs) [5, 6]. In this Section, we provide a comprehensive study of the literature review showing contributions to the state-of-the-art in which experts employ AI techniques in the legal field.
In [7], researchers presented a new model named ITALIAN LEGAL-BERT, which can interpret the semantic nuances of Italian legal texts with unprecedented accuracy. In detail, they proposed two versions of that model using two different methods of domain adaptation: one based on Italian civil cases and another on Italian legal documents based on the CamemBERT architecture. Different from them, in our experiment, we proposed the usage of LLMs to verify the compliance of online privacy policies, using a small dataset manually validated by a legal expert. Once the validation phase was concluded, we used the dataset to fine-tune five different models belonging to state-of-the-art. After the appearance of LLMs the approach changed significantly, at least partially abandoning the BERT architectures.
Rodriguez *et al.* in [8] presented a study in which the researchers employed LLMs to analyze privacy policies by automating the extraction and interpretation of critical information at scale. In detail, the authors reported performances obtained using ChatGPT and Llama 2 *i.e.*, popular LLM-based tools. At the end of the experiments, they achieved interesting results. Differently from our work, researchers employed four different datasets (MAPP, OPP-115, APP-350, and IT-100) belonging to several fields. However, in our study we focused only on the Policy Privacy of websites with the European domain, in English. The privacy policies were selected by the criterion that they were typical in the content and style of phrasing of privacy policies found online so that they would be a better representative sample. The text of online privacy policies tend to use standard phrases, and therefore the samples which

were selected are consistent with the numerous policies that were read for this exercise. Moreover, we employed a large set of LLMs such as BERT-base-uncased, roBERTa-base, distilBERT-base-uncased, t5-base, and ERNIE-2.0-base-en.
Garza *et al.* in [9] designed a novel approach named PrivComp-KG to manage and verify privacy policy compliance efficiently. In detail, they leveraged LLMs and Retrieval Augmented Generation (RAG) to identify relevant privacy policy sections and match them with corresponding legal rules, achieving high score accuracy. On the other hand, in our proposed method, instead of using an open-science dataset, we created a dataset from scratch, downloading several privacy policies from the web (by using a Python script written by the authors) and then manually validated by the legal expert in our research team, and we selected a specific clause of the GDPR Article 13(2)(b), which is the duty to inform the data subject of the right to rectification and erasure of personal data.

3. The Method

This Section provides an overview of the proposed method to classify a privacy policy that complies with Art.13(2)(b) of the GDPR. Figure 1 illustrates the steps performed in detail.

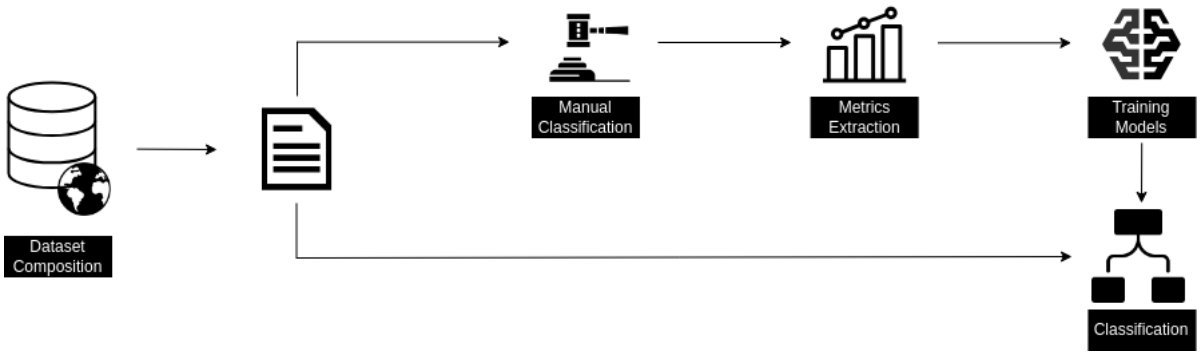

Figure 1: The proposed method to classify privacy policy complies with Art.13(2)(b) of the GDPR.

3.1. Dataset Composition

To automate evaluating whether or not a given privacy policy complies with Art.13(2)(b) of the GDPR as a preliminary stage, we developed a Python script to retrieve the Policy Privacy page of several websites located within the EU. In detail, we defined the privacy policy slugs list with several slugs typically employed in websites to store privacy policies like */privacy /privacy-policy, /privacy-notice, /privacy-statement, /legal/privacy-policy, /legal/privacy, /policies/privacy, /about/privacy, /about-us/privacy, /info/privacy, /support/privacy, /help/privacy, /terms/privacy, /resources/privacy,* end */company/privacy.* Once we identified these websites, considering only websites with European domains, we employed another Python script developed by the authors to extract the text from the retrieved URLs. The script was designed to clean and format the extracted content, removing webpage elements such as buttons and other non-text components. In detail, we considered only privacy policies in English, which were divided into sentences, removing stop words and punctuation marks.

3.2. Manual Classification

After composing the dataset and converting all of the privacy policies into text files, we moved on to the next stage of the experiment: the manual classification phase to ensure a qualitative dataset. Human annotators can guarantee that labels are accurate and consistent, which is crucial for training effective machine learning models, especially in a supervised learning task as delicate as legal compliance. Additionally, this process can provide transparency and help others understand how the dataset was created. Given that this type of task is time-consuming and resource-intensive, we decided to take a

subset of privacy policies retrieved (10 out of 50). Due to the small amount of data (less than 30 samples), we calculated the confidence interval using the t-distribution. The calculated confidence interval (5.91,8.29) represents the range within which we are 95% confident the true mean of the population lies. This interval is based on the sample mean (7.1) and the sample standard deviation (1.66), statistical estimates which were derived from our data. The critical t-value (2.262) at a 95% confidence level sets the margin of error, reflecting how much we expect the sample mean to vary from the population mean. The resulting interval captures the inherent uncertainty in estimating population parameters from a small sample and provides an actionable range for making informed decisions with high confidence. Once the number of samples was identified, a legal expert in the group manual classified the ten selected privacy policies. Regarding the manual classification phase, the first challenge, then, was to select a suitable article from the GDPR to test the LLM and determine if its requirements were satisfied within the text of online privacy policies. Article 13, "Information to be provided where personal data are collected from the data subject" and Article 14, "Information to be provided where personal data have not been obtained from the data subject" were the best choices in terms of checking for website information because these two articles list the information that must be supplied to the data subject. Informing data subjects of their rights (as part of the principle of transparency) is demonstrated by using clear and plain language provided in an easy display, in a place where the data subject is likely to find it (GDPR Rec. 39). More specifically, the clause 13(2)(b) "the existence of the right to request from the controller access to and rectification or erasure of personal data" is a duty to inform of this right that could be demonstrated as compliant on its face (meaning in its language alone). Essentially then this means being compliant with the applicable articles of the GDPR that would be required to be present in a privacy policy on a website that intends to gather or process personal data. Several sample privacy policies that were viewed for this project were determined to be a mix of compliant and non-compliant wording. A manual selection of privacy policies on websites were made from the criteria that they were 1.) in English, and 2.) in the EU. Other parameters that may help demonstrate "compliance" such as location on the website and size of font, are outside the scope of this research project. The selection of privacy policies would serve as one small test scenario on which the method could be tested, and further expanded based on positive results, or at least adjusted and measured against reliable results.

3.3. Metrics Extraction

The dataset composition is essential to training a model. For this reason, we decided to use a larger dataset that provides more examples for the model to learn from, which can improve its ability to generalize to unseen data [10]. With only this data (those retrieved in Section 3.2), models may memorize the training examples rather than learning the underlying patterns, leading to overfitting. A larger dataset helps mitigate this risk by providing more diverse examples. Also, a larger dataset is more likely to capture a wide variety of scenarios, edge cases, and variations in the data. This diversity helps the model learn to handle different situations and improves its robustness. To speed up the process and increase the dataset size, this work aims to carry out a semantic sentence similarity process. The goal is to rely on Sentence Transformers, like [11], and LLMs like [12, 13, 14] to obtain contextual embeddings from the labeled sentences and calculate their similarity with Art. 13(2)(b). Once we calculated this similarity, we used this score as a basis to label new unseen sentences from other privacy policies. The assumption is that the calculated score between the manually labeled sentences and Art. 13(2)(b) can be used as a threshold to evaluate the compliance of new sentences. Sentence transformers are models designed to convert sentences into fixed-size vector representations, capturing their semantic meaning. These models are used for tasks like semantic similarity, clustering, and information retrieval, demonstrating excellent results [18]. On the other hand, LLMs are advanced neural networks trained on vast amounts of text data to understand and generate human-like text, enabling them to perform a wide range of language-related tasks, demonstrating excellent results for this purpose [15, 16, 17]. Regarding the similarity score, we used three metrics commonly used in Natural Language Processing (NLP): cosine similarity, dot product, and Euclidean distance. Cosine similarity measures the cosine of the angle between two non-zero vectors in a multi-dimensional space. It is beneficial for measuring

the similarity of text data represented as vectors because it focuses on the orientation of the vectors rather than their magnitude, making it practical for comparing documents of different lengths. The dot product of two vectors is a scalar value, the sum of the products of their corresponding components. It can indicate the degree of similarity between two vectors. A higher dot product suggests greater similarity, but it is sensitive to the magnitude of the vectors, which can be a limitation when comparing vectors of different lengths, which does not affect the case of Sentence Transformers since they produce fixed-size vector representations. At the same time, the Euclidean distance measures the straight-line distance between two points (or vectors) in Euclidean space. It provides a measure of the absolute distance between two vectors. NLP can be used to assess how similar or dissimilar two sentences are based on their vector representations. Using the score obtained with the previously described metrics, between each manually labeled sentence and Art.13(2)(b), we aimed to label a new unseen privacy policy sentence as compliant or not if this score is equal to or greater.

Text	Link	Compliant
Right to rectification and erasure: If your data is incorrect or incomplete, you can request us to amend or supplement it. You can also request that your personal data be deleted. We strive to process your request as quickly as possible, but at the latest within four weeks. Please note that we may not be able to delete all your data because we are bound by certain laws, such as the fiscal retention obligation.	https://www.twence.com/privacy-statement	Yes
Esri will permit you to access, correct, or delete your information in our database by contacting accounts@esri.com or by logging in to your account at https://my.esri.com/ and making the appropriate changes or selecting the “Remove my account” option. We will respond to all requests for access within a reasonable timeframe.	https://www.esri.com/en-us/privacy/privacy-statements/privacy-statement	Yes

Table 1
Example of a sentence of a privacy policy compliant with Art.13(2)(b) of the GDPR with the corresponding link.

3.4. Experiments

As explained, we obtained the sentence embeddings from both Sentence Transformers and Large Language Models to perform the contextual sentence similarity task. Regarding the first type, we selected the five best models based on their average performance on encoding sentences over 14 diverse tasks from different domains[1], reported in Table 2. Regarding the LLMs, we selected the BERT model [12], which is the first bi-directional (or non-directional) pre-trained language model (BERT-base-uncased), and two different variants of the latter that are DistilBERT [18] (distilBERT-base-uncased) aiming to optimize BERT’s performance and RoBERTa [19] (roBERTa-base), short for "Robustly Optimized BERT Approach." In addition, we employed the T5 model (t5-base), or Text-to-Text Transfer Transformer, a large language model [20] developed by Google AI that uses a text-to-text approach. Finally, ERNIE [21], Enhanced Representation through kNowledge IntEgration (ERNIE-2.0-base-en) consists of two stacked modules: a textual encoder and a knowledgeable encoder, which is responsible for integrating extra token-oriented knowledge information into textual information.

[1] https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#original-models

Model	Description
all-mpnet-base-v2	All-round model tuned for many use-cases. Trained on a large and diverse dataset of over 1 billion training pairs
multi-qa-mpnet-base-dot-v1	This model was tuned for semantic search: Given a query/question, it can find relevant passages. It was trained on a large and diverse set of (question, answer) pairs
all-distilroberta-v1	All-round model tuned for many use-cases. Trained on a large and diverse dataset of over 1 billion training pairs
all-MiniLM-L12-v2	All-round model tuned for many use-cases. Trained on a large and diverse dataset of over 1 billion training pairs
multi-qa-distilbert-cos-v1	This model was tuned for semantic search: Given a query/question, it can find relevant passages. It was trained on a large and diverse set of (question, answer) pairs

Table 2
Sentence Transformers models involved in the task of contextual sentence similarity.

4. Results

As described in the previous section, we used the ten manual classified sentence policies and calculated their similarity scores based on GDPR Art.13(2)(b) right to rectification and erasure. Table 3 reports the average score metrics for each model. We used these scores as a threshold to classify new unseen privacy policy sentences automatically. In order to test this approach, we randomly extracted a privacy policy from the dataset created in Section 3.1. We split it into sentences, submitted them to each model, and calculated the metrics described in Section 3.3, *i.e.*, cosine similarity, dot product, and Euclidean distance against Art.13(2)(b). After this process, we extracted 168 privacy policy. Whenever one of the calculated metrics satisfied the previously calculated threshold, we labeled the sentence as compliant.

Sentence Transformers Models			
Model	**CSM**	**DPM**	**EDM**
all-mpnet-base-v2	0.445	0.445	-1.041
multi-qa-mpnet-base-dot-v1	0.530	22.509	-6.217
all-distilroberta-v1	0.462	0.462	-1.024
all-MiniLM-L12-v2	0.491	0.491	0.999
multi-qa-distilbert-cos-v1	0.503	0.503	-0.985
Large Language Models			
Model	**CSM**	**DPM**	**EDM**
BERT-base-uncased	0.794	67.855	-5.874
roBERTa-base	0.971	160.567	-2.998
distilBERT-base-uncased	0.857	59.430	-4.374
t5-base	0.729	7.430	-2.373
ERNIE-2.0-base-en	0.888	88.184	-4.593

Table 3
Models type, the related cosine similarity mean score, dot product mean score, and Euclidean distance mean score were calculated between the manually labeled compliance policy sentence and Art.13(2)(b).

LEGEND:
CSM: Cosine Similarity Mean **DPM:** Dot Product Mean **EDM:** Euclidean Distance Mean

Table 4 shows the results of this phase after the predictions extracted by the models were manually checked. The Table shows the false positive results (where the sentence was predicted as compliant but were not) and the false negatives (*i.e.*, the sentence was predicted as not compliant but actually was compliant). While certain models exhibited a high success rate, some failed due to high false positive occurrences. Among the models tested, multi-qa-distilBERT-cos-v1 and all-MiniLM-L12-v2 were the best-performing models, as each model produced two false positives and one false negative. This implies

Sentence Transformers Models			
Model	**False Positives**	**False Negatives**	**Total Sentences**
all-mpnet-base-v2	18	1	168
multi-qa-mpnet-base-dot-v1	11	1	168
all-distilroberta-v1	4	1	168
all-MiniLM-L12-v2	2	1	168
multi-qa-distilbert-cos-v1	2	1	168
Large Language Models			
Model	**False Positives**	**False Negatives**	**Total Sentences**
BERT-base-uncased	41	1	168
roBERTa-base	55	2	168
distilBERT-base-uncased	44	1	168
t5-base	31	0	168
ERNIE-2.0-base-en	64	0	168

Table 4
Model types and the number of false positives and false negatives privacy policy sentences after the manual check.

that these models possess a vague but dependable general semantic understanding with no adjustment or fine-tuning for legal text. The roBERTa-base and distilBERT-base-uncased models predicted many non-compliant sentences as compliant. It means roBERTa-base misclassified 55 sentences, attributing them to the compliant position because of a tendency to superficially overlap with the semantics of those in the compliant position. Since all models used in this experiment were pre-trained and not fine-tuned on GDPR-specific datasets, their performance reflects their general-purpose capabilities rather than specialized legal expertise. Therefore, this means that the models were not legal experts in any particular area but were general models. Indeed, the pretrained models were able to assist in providing information about the importance of the semantic similarity of privacy policy sentences to the requirements of GDPR.

5. Future Work

In this preliminary experiment, we designed a methodology to identify phrases within websites' European English language privacy policies, so that these phrases could be used to determine if they fulfill the requirement imposed by the GDPR on data controllers to inform the data subject of the right to rectify or erase personal data collected from them (Art. 13(2)(b)) by leveraging LLMs. As a first step, we employed a Python script written by authors to retrieve websites' policies and create a preliminary dataset. Next, we manually classified ten privacy policies and then we calculated several metrics like Cosine Similarity Mean, Dot Product Mean, and Euclidean Distance Mean leveraging pre-trained models. We tested the model using a single privacy policy that was not included in the training phase.
To move our research forward, we will refine our model using a Small Language Model (SML) with cosine similarity as the evaluation metric. As a matter also of future work, a concern to be addressed is the risk of a research design that leads to a self-fulfilling prophecy: Since many privacy policies in their information sheets "mirror" the text of the GDPR, it is possible that the similarity score returns positive results from a semantic point of view without reflecting actual compliance. Other next steps could be to expand to other GDPR requirements that make up a typical privacy policy. Additionally, we will expand our dataset by retrieving a diverse set of internet privacy policies. By leveraging few-shot learning techniques, we strive to achieve high accuracy with limited data, optimizing our model's ability to generalize effectively to new examples.

Acknowledgments

This work has been partially supported by EU DUCA, EU CyberSecPro, SYNAPSE, PTR 22-24 P2.01 (Cybersecurity) and SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the EU - NextGenerationEU projects, by MUR - REASONING: foRmal mEthods for computAtional analySis for diagnOsis and progNosis in imagING - PRIN, e-DAI (Digital ecosystem for integrated analysis of heterogeneous health data related to high-impact diseases: innovative model of care and research), Health Operational Plan, FSC 2014-2020, PRIN-MUR-Ministry of Health, the National Plan for NRRP Complementary Investments D^3 4 Health: Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care, Progetto MolisCTe, Ministero delle Imprese e del Made in Italy, Italy, CUP: D33B22000060001, FORESEEN: FORmal mEthodS for attack dEtEction in autonomous driviNg systems CUP N.P2022WYAEW and ALOHA: a framework for monitoring the physical and psychological health status of the Worker through Object detection and federated machine learning, Call for Collaborative Research BRiC -2024, INAIL - NextGenerationEU projects, SMAUG, Horizon Europe GA number 101121129, Fit4MedRob: This work was supported by the Italian Ministry of Research, under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (# PNC0000007)

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2024. Online; accessed 14 Dec 2024.

[2] https://huggingface.co/sentence-transformers/all-distilroberta-v1, 2024. Online; accessed 12 Dec 2024.

[3] F. Ariai, G. Demartini, Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges, arXiv preprint arXiv:2410.21306 (2024).

[4] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, M. J. Bommarito II, Natural language processing in the legal domain, arXiv preprint arXiv:2302.12039 (2023).

[5] J. Lai, W. Gan, J. Wu, Z. Qi, S. Y. Philip, Large language models in law: A survey, AI Open (2024).

[6] D. H. Anh, D.-T. Do, V. Tran, N. Le Minh, The impact of large language modeling on natural language processing in legal texts: a comprehensive survey, in: 2023 15th International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2023, pp. 1–7.

[7] D. Licari, G. Comandè, Italian-legal-bert models for improving natural language processing tasks in the italian legal domain, Computer Law & Security Review 52 (2024) 105908.

[8] D. Rodriguez, I. Yang, J. M. Del Alamo, N. Sadeh, Large language models: a new approach for privacy policy analysis at scale, Computing 106 (2024) 3879–3903.

[9] L. Garza, L. Elluri, A. Kotal, A. Piplai, D. Gupta, A. Joshi, Privcomp-kg: Leveraging knowledge graph and large language models for privacy policy compliance verification, arXiv preprint arXiv:2404.19744 (2024).

[10] D. Rajput, W.-J. Wang, C.-C. Chen, Evaluation of a decided sample size in machine learning applications, BMC bioinformatics 24 (2023) 48.

[11] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[12] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[13] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, L. Okruszek,

Detecting formal thought disorder by deep contextualized word representations, Psychiatry Research 304 (2021) 114135.

[14] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, CoRR abs/1906.08237 (2019). URL: http://arxiv.org/abs/1906.08237. arXiv:1906.08237.

[15] M. Ormerod, J. Martínez del Rincón, B. Devereux, Predicting semantic similarity between clinical sentence pairs using transformer models: Evaluation and representational analysis, JMIR Medical Informatics 9 (2021) e23099.

[16] M. Freestone, S. K. K. Santu, Word embeddings revisited: Do llms offer something new?, arXiv preprint arXiv:2402.11094 (2024).

[17] M. T. R. Laskar, X. Huang, E. Hoque, Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 5505–5514.

[18] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[21] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, H. Wu, Ernie: Enhanced representation through knowledge integration, arXiv preprint arXiv:1904.09223 (2019).