NLP for GDPR Compliance Checking

Ozioma Okonicha

Table of contents

Table of contents

Introduction

Natural Language Processing (NLP)

is a field of artificial intelligence that focuses on the interaction between computers and humans using natural language.

General Data Protection Regulation (GDPR)

is a comprehensive data protection law in the European Union, governing the processing of personal data.

Introduction

Log

From searching primary sources like Google scholar

33

20

Read

out of the 33 papers added to the log

Compliance and GDPR Audit

Compliance

Compliance refers to the adherence of an organization's practices, policies, and processes to relevant laws and regulations.

GDPR Compliance Audit

A GDPR compliance audit involves assessing and verifying that an organization follows the principles and requirements outlined in the GDPR.

Relevance of GDPR

GDPR addresses issues like:

01

Data protection

02

Consent

03

User rights

requiring organizations to adopt transparent and lawful practices.

Background

01

Context

- Importance of personal data and aligning processing agreements with GDPR regulations.
- Complexity of GDPR requirements and the need for compliance audits.

Background

02

Problem

- Resource-intensive manual auditing of Data Processing Agreements (DPAs) and Privacy Policy Statements.
- Manual audits are resource-consuming and prone to errors.
- Attempts to automate compliance checking face limitations in accuracy and coverage.

Background

03

Motivation

- GDPR enforcement includes heavy fines for non-compliance.
- Automated compliance checking with NLP can mitigate legal risks and fines.

Background

04

Challenges

- Complexity of GDPR regulations, requiring in-depth understanding.
- Ambiguity and verbosity in legal texts, often necessitating human expertise.
- Papers in the literature focus on specific aspects of GDPR, not comprehensive audits.

Tasks for NLP in GDPR

01	**Privacy Policy Analysis:**	Understanding and simplifying privacy policies. [1, 11, 12]
02	**Compliance Checking:**	Verifying if policies align with GDPR regulations. [2, 4, 7, 11, 14, 16, 17]
03	**Semantic Annotation**	Enabling semantic search over legal texts. [6, 18]

Tasks for NLP in GDPR

01

Privacy Policy Analysis:

Recommend policies for specific issues based on past effectiveness and relevance.

Named Entity Recognition (NER): Utilizing NER to identify and classify GDPR-related entities in privacy policies,

Text Classification

Tasks for NLP in GDPR

02 **Compliance Checking:**

Automatically identify non-compliance in contracts and legal documents.

Semantic Similarity Analysis to understand the context and meaning of text segments related to GDPR compliance.

Developing an NLP model to compare organizational privacy policies against GDPR requirements, identifying gaps or areas of non-compliance.

Tasks for NLP in GDPR

03 Semantic Annotation

Legal Entity Recognition: Identify and classify legal entities in text documents.

Semantic Role Labeling in Policy Documents: Analyze the semantic roles within sentences to understand policy document structures.

Relation Extraction

Key Insights

02

Models

Various models, including Logistic Regression, SVM, and BERT, are employed for compliance checking. [2, 4, 7, 8, 11, 16, 17]

04

Metrics

Evaluation metrics include F1 scores, precision, and recall. [2, 4, 7, 11, 16, 17]

01

Datasets

Datasets like OPP-115, PrivaSeer, and GDPR excerpts are used for evaluation. [1, 2, 4, 6, 7, 8, 10, 11, 12, 16, 17, 18]

03

Limitations

Limitations often include small datasets, reliance on expert knowledge, and challenges in generalizability. [2, 4, 7, 8, 11, 16, 17]

Trends and Patterns

Application of Machine Learning Models

[2, 4, 6, 8, 11, 15, 20]

Use of NLP Techniques

[2, 6, 8, 11, 12, 15, 18, 19, 20]

Automated Compliance Checking

[2, 6, 7, 11, 16, 17

Focus on Specific Aspects of GDPR

[1, 4, 7, 10, 11, 13, 17]

Trends and Patterns

Dataset Creation

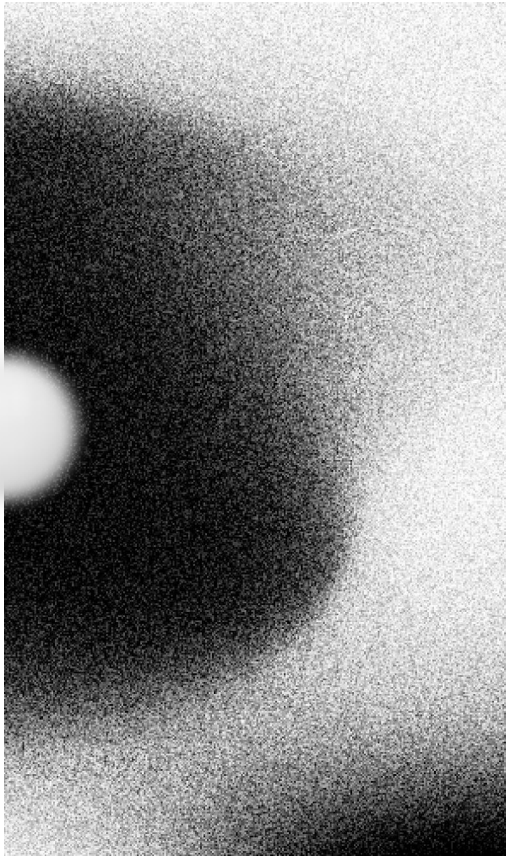
Several papers contribute by creating datasets specific to privacy policies.

[1, 4, 7, 12, 15, 20]

Interdisciplinary Approaches

Many papers combine legal expertise with NLP techniques

[1, 6, 13, 17, 18]



Lessons Learned

01

Role

NLP plays a crucial role in automating GDPR compliance tasks.

02

Challenges

Challenges remain, such as the need for larger and diverse datasets.

Research Questions

RQ1

How can NLP models contribute to automating the process of compliance checking with data protection laws, reducing manual efforts?

RQ2

How do NLP models adapt to changes in privacy policies, especially considering the dynamic nature of legal documents that may undergo updates or amendments?

Research Questions

RQ?

How resource-efficient are NLP-based compliance checking systems, particularly in terms of computational requirements and processing time?

Proposed Method

01

Explore pre-trained models

Check the recent pre-trained models, even beyond BERT. Maybe GPT-3 and T5, how can they be applied to GDPR compliance checking?

02

Perform fine-tuning

Develop a fine-tuning strategy that will specifically target GDPR nuances. Maybe transfer learning can be used to adapt existing models to the legal language used in privacy policies.

Model Selection

Option 1

How useful is GPT-3 in understanding and generating legal language?

Can it be useful or even efficient in potential for compliance checking tasks?

GPT-3

T5

Option 2

Can T5 be used as a text-to-text model for handling legal text?

Can it transform privacy policies into compliance related representations?

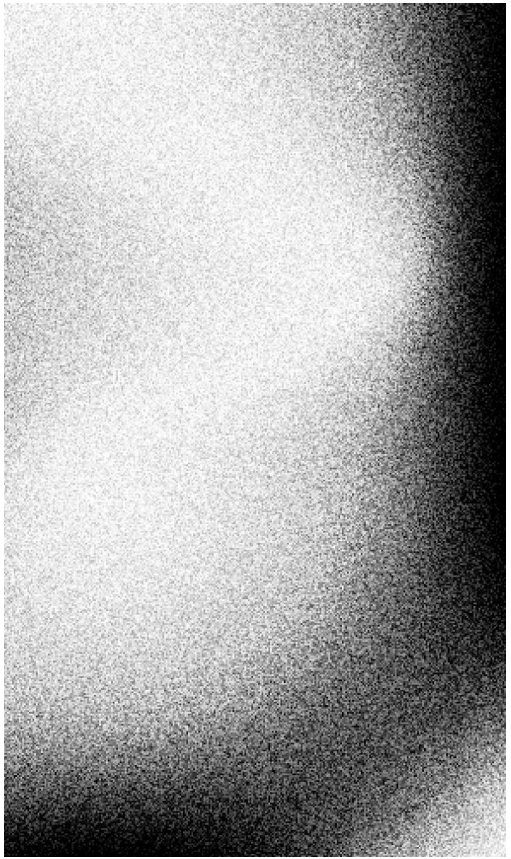
Experiment Strategy

01. Pre-training and Fine-tuning

- **Model Selection**: Explore pre-trained models like GPT-3 and T5 alongside BERT, evaluating their performance on GDPR-specific tasks (e.g., classification, information extraction).

- **Dataset Preparation**: Create a comprehensive dataset of privacy policies annotated with GDPR provisions and compliance flags.

- **Fine-tuning:** Adapt chosen models to the legal domain by fine-tuning on the GDPR dataset, focusing on accuracy and interpretability.



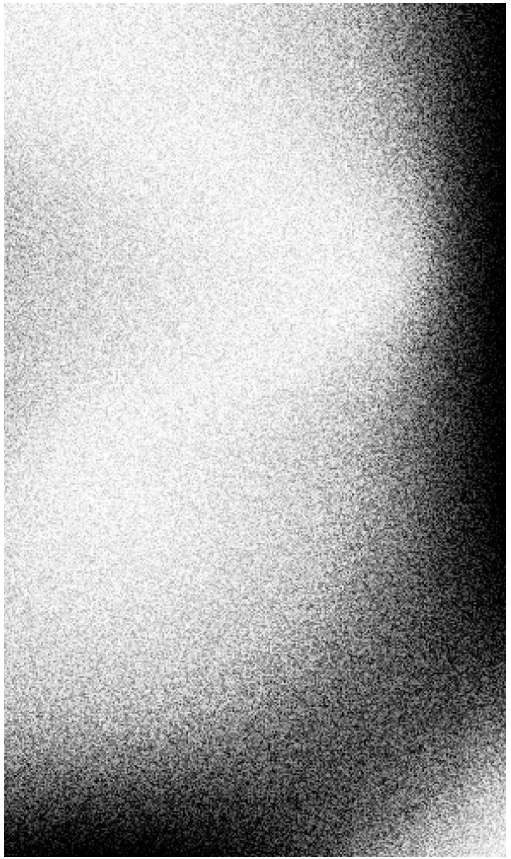
Experiment Strategy

02. Evaluation and Refinement

- **Metrics:** Use several metrics like precision, recall, F1-score, positives/negatives to understand the numerical performance of the model.

- **Human evaluation:** Conduct human evaluations of model outputs for accuracy and interpretability, refining the model based on feedback.

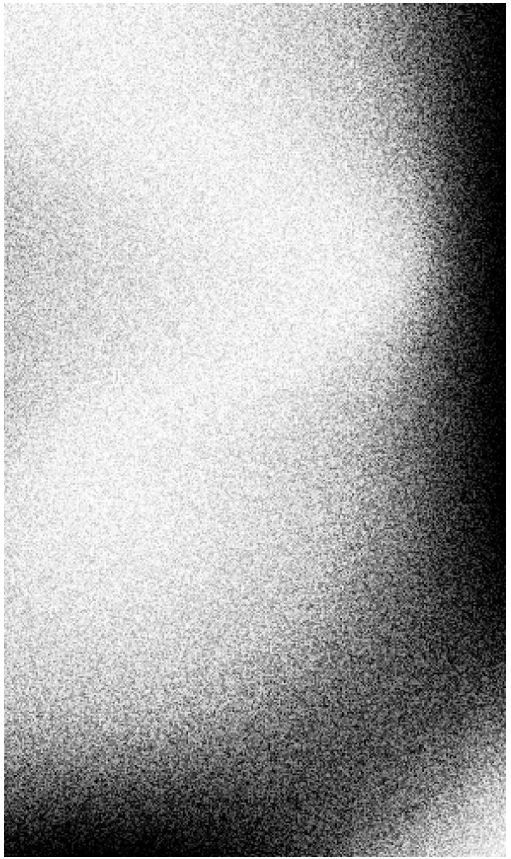
- **Model comparison:** Compare the performance of different models, evaluate them and identify the optimal choice that answers the research questions.



Experiment Strategy

03. Adaptation and Monitoring

- **Continuous learning:** Implement periodic retraining on updated privacy policies and legal rulings to maintain model accuracy and adaptability.
- **Active learning:** Integrate mechanisms for querying human experts on ambiguous cases or emerging legal interpretations.
- **Performance monitoring:** Track the model's performance in real-world settings and adjust the training data or fine-tuning approach as needed.



Evaluation Questions

Compliance elements

How accurately does the NLP model identify and extract key compliance elements from privacy policies?

Sensitivity

How sensitive is the NLP model to updates or amendments in privacy policies, and how quickly does it adapt to these changes?

Legal details

How does the selected model handle legal language intricacies present in GDPR and other data protection laws?

Fine-tuning

What impact does fine-tuning have on model performance, and how well does it adapt to changes in privacy policies?

Results

Metric/Model	BERT (OPP Dataset)	SBERT	GPT-2
Dataset	OPP (115 Policies)	ACL Coling	ACL Coling
Training Loss	Gradually Decreased	-	-
Validation Loss	Decreased to 0.298	-	-
Exact Match Ratio (EMR)	Increased to 0.652	-	-
F1 Score (Samples)	Increased to 0.953	-	-
Cosine Similarity Threshold	0.9	0.5	0.6
Number Passed Threshold	273 out of 10501	95	10499

Thanks!