UNIVERSITÄT PASSAU

OSSYM 2024

A DATASET OF GDPR COMPLIANT NER FOR PRIVACY POLICIES

HARSHIL DARJI, STEFAN BECHER, JELENA MITROVIC, ARMIN GERL, MICHAEL GRANITZER

UNIVERSITY OF PASSAU, GERMANY

AGENDA

1. Introduction
2. Background
    1. Privacy policy languages
    2. Privacy preference languages
3. Related Work
4. Dataset Overview
5. Inter-Annotator Agreement
6. Fine-Tuning Language Models
    1. Confusion matrix
7. Conclusion
8. Future work

1. INTRODUCTION

- Privacy policies inform users about data practices.
- Policies are lengthy and complex, confusing users.
- NER helps extract information from policies.
- We present a dataset of NER-annotated privacy policies.
- The dataset includes privacy policies from 44 different online platforms.
- Policies manually annotated for GDPR compliance.
- Bridges the gap between GDPR requirements and real-life policy implementation, improving accessibility and understanding.

2. BACKGROUND

- GDPR (General Data Protection Regulation) is a comprehensive data protection law in the European Union.
- It protects the privacy and personal data of individuals within the EU by setting high standards for data handling.
- It ensures transparency and accountability in how organizations manage personal data.
- It applies to all organizations that process the personal data of EU citizens.
- It introduces strict requirements for obtaining consent from individuals whose data is being collected.
- Regulations mandate that personal data must be collected for specific, explicit, and legitimate purposes only.

2.1. PRIVACY POLICY LANGUAGES

- Classical privacy policies are represented as huge plain-text documents only, which are hardly understandable.
- Privacy policy languages (e.g., SPECIAL[1], LPL[2]) add a technical structure to these classical privacy policies.
- Therefore, privacy policies become processable and can be presented in a user-friendly format.
- Advantages are that policy statements can be technically enforced and policy content is easier accessible.
- Annotations can be used to automatically translate plain-text privacy policies if no technical format is given.

[1] Scalable policy-aware linked data ar-chitecture for privacy, transparency and compliance.
[2] Gerl, A., Bennani, N., Kosch, H., and Brunie,L. (2018). Lpl, towards a gdpr-compliantprivacy language: formal definition and us-age. In Transactions on Large-Scale Data-andKnowledge-Centered Systems XXXVII, pages 41–80.Springer.

2.2. PRIVACY PREFERENCE LANGUAGES

- Fully reading/understanding privacy policies takes an unreasonable amount of time (10 – 20 min per policy) [1].
- Privacy preference languages (e.g., ConTra[2], YaPPL[3]) let users define their preferences regarding privacy policy statements.
- Preferences and privacy policies are automatically matched, and users get feedback about mismatches.
- Therefore, only mismatching parts of the privacy policy have to be checked manually, which shortens the required amount of time to understand privacy policies by a lot.
- Privacy policies are required to be defined in a technical format for this process.
- Again, annotations can be used for an automated translation.

[1] Ibdah, D., Lachtar, N., Raparthi, S. M., and Bacha,A. (2021). "why should i read the privacy policy,i just need the service": A study on attitudes andperceptions toward privacy policies. IEEE Access,9:166465–166487
[2] Becher, S. and Gerl, A. (2022). Contra preference lan-guage: Privacy preference unification via privacy in-terfaces. Sensors, 22(14)..
[3] Ulbricht, M.-R. and Pallas, F. (2018). Yappl - alightweight privacy preference language for legallysufficient and automated consent provision in iot sce-narios. In DPM/CBT@ESORICS, pages 329–344,09.

3. RELATED WORK

- Previous studies have focused on the automated analysis of privacy policies using various techniques.[1]
- Research has identified the potential for extracting individual pieces of information from privacy policies to enhance understanding.
- PrivacyGLUE benchmark has been developed for evaluating language models specifically in the context of privacy policies.[2]
- Existing datasets like OPP-115 and APP-350 have limitations, such as being outdated or not fully GDPR-compliant.
- The OPP-115 dataset contains 115 manually annotated privacy policies created before GDPR implementation.[3]
- The APP-350 dataset includes 350 mobile app privacy policies annotated for compliance issues but is limited to mobile apps.[4]
- A large corpus of over one million privacy policies has been collected, but it lacks NER annotations necessary for detailed analysis.[5]

[1] J. M. Del Alamo, D. S. Guaman, B. García, and A. Diez, "A systematic mapping study on automated analysis of privacy policies," Computing, vol. 104, no. 9, pp. 2053–2076, 2022.
[2] A. Shankar, A. Waldis, C. Bless, M. Andueza Rodriguez, and L. Mazzola, "Privacyglue: A benchmark dataset for general language understanding in privacy policies," Applied Sciences, vol. 13, no. 6, p. 3701, 2023.
[3] S. Wilson et al., "The creation and analysis of a website privacy policy corpus," 2016, pp. 1330–1340. 10.18653/ v1/P16-1126
[4] S. Zimmeck et al., "Maps: Scaling privacy compliance analysis to a million apps," Proceedings on Privacy Enhancing Technologies, vol. 2019, pp. 66–86, 2019. 10.2478/ popets-2019-0037
[5] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. R. Mayer, "Privacy policies over time: Curation and analysis of a million-document dataset," CoRR, vol. abs/2008.09159, 2020. https: //arxiv.org/abs/2008.09159

4. DATASET OVERVIEW

- The dataset consists of 44 European privacy policies, manually annotated by legal experts.
- Annotations are based on the Data Privacy Vocabulary (DPV) to ensure GDPR compliance.
- DPV includes categories like Data Controller (DC), Data Processor (DP), Data Protection Officer (DPO), and more.
- The label set includes 33 categories such as Data Subject, Data Source, Processing, Personal Data, and Non-Personal Data.
- Key GDPR-related legal terms annotated include Organisational Measure (OM), Technical Measure (TM), Legal Basis (LB), and Consent (CONS).
- Annotations also cover specific GDPR Data Subject Rights like the Right to Access (Art. 15), Right to Rectification (Art. 16), and Right to Erasure (Art. 17).
- The dataset follows the CoNLL-2002 format and is publicly available for research and practical applications.

4. DATASET OVERVIEW

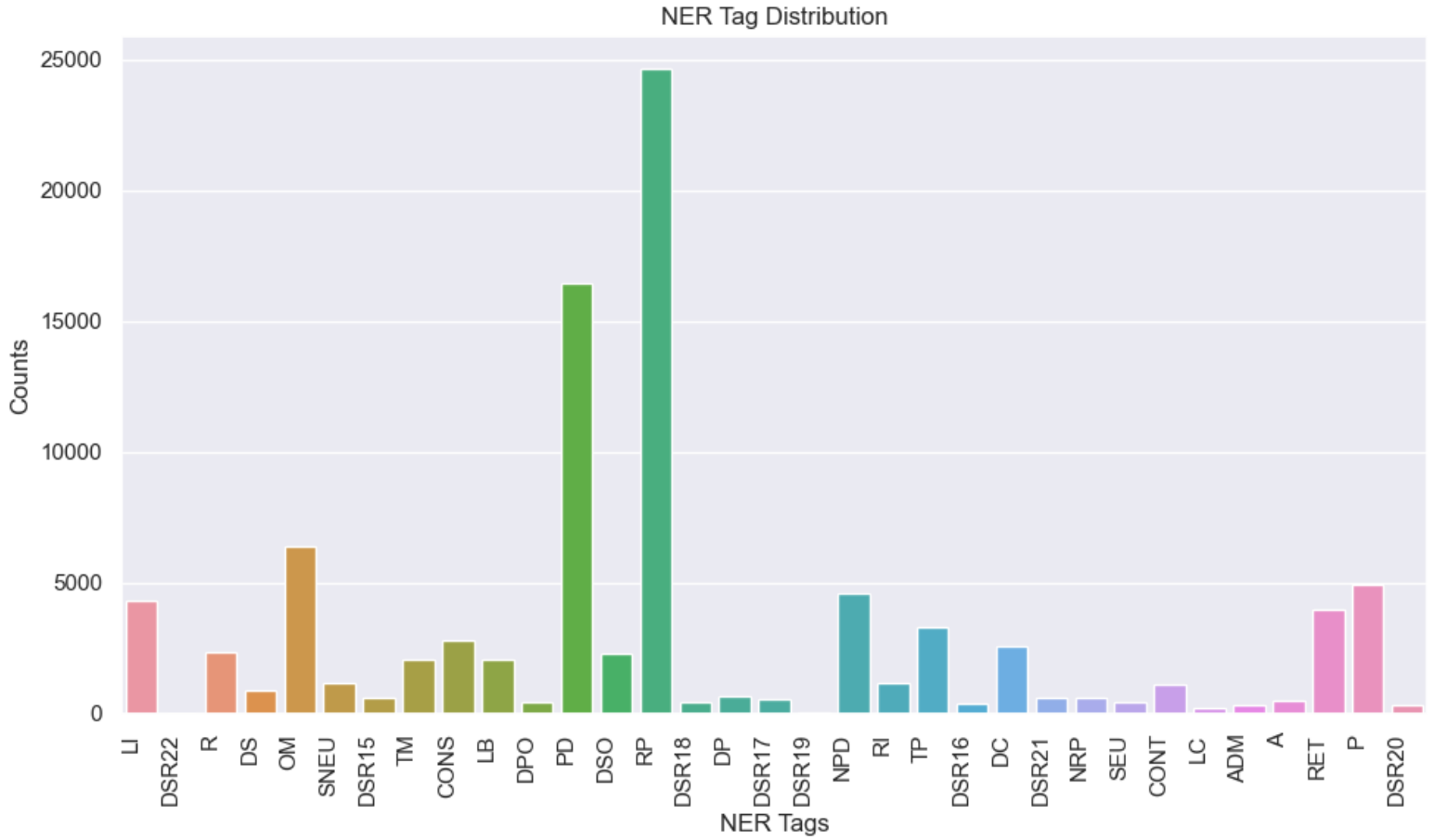

*Figure **1**. The number of occurrences of each NER tag in the annotated data set.*

5. INTER-ANNOTATOR AGREEMENT

- The Cohen's Kappa score between two annotators on 20 documents was relatively low at 0.6412.
- Upon careful, this low score was not caused by disagreement in labelling.
- Annotators used Word’s comment feature for annotation, leading to inconsistencies and discrepancies in how the text was selected and annotated.
- These minor differences, while seemingly trivial, can affect automated processing.
- This impacts agreement scores, making it seem like annotators disagree when, in reality, they align in understanding but differ in text selection.

6. FINE-TUNING LANGUAGE MODELS

- The dataset was used to fine-tune five language models: BERT (*bert-base-cased*), GPT-2, ALBERT (*albert-base-v2*), RoBERTa (*roberta-base*), and BERT multilingual (*bert-base-multilingual-cased*).
- The aim of this experiment was to identify the best model for recognizing Named Entities (NER) in complex privacy policy texts.
- BERT (*bert-base-cased*) distinctly outperforms its counterparts, registering an F1-score of 0.74.
- Results indicate the usefulness of the GDPR-compliant dataset in enhancing the capabilities of language models for privacy policy analysis.

Language model	F1-score
bert-base-cased	0.74122
gpt2	0.71343
albert-base-v2	0.71192
roberta-base	0.72689
bert-base-multilingual-cased	0.73534

6.1. CONFUSION MATRIX

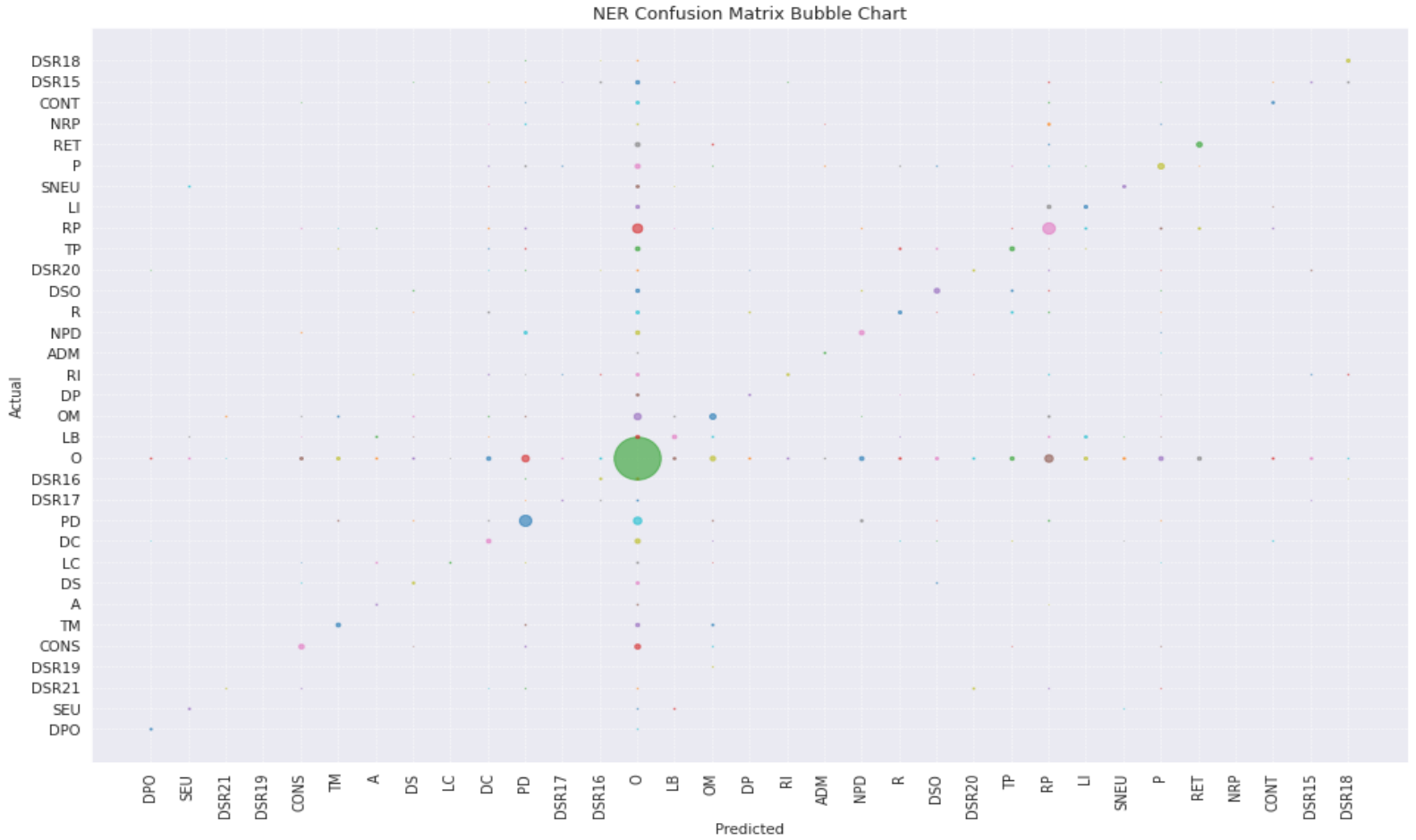

Figure 2. Confusion matrix for bert-base-cased. Each cell in this confusion matrix represents the count of instances where a particular entity type (as represented by rows) is predicted as another type (as represented by columns). It's important to note that the "O" tag, representing words that are not part of any named entity, dominates the confusion matrix due to its high occurrence in the dataset.

7. CONCLUSION

- Presents a GDPR-compliant dataset of 44 privacy policies from European platforms.
- Annotated with Named Entity Recognition (NER) tags by legal experts.
- Annotations follow the Data Privacy Vocabulary (DPV) and cover 33 GDPR-relevant categories.
- Fine-tuning five language models revealed BERT (bert-base-cased) achieved the highest F1-score of 0.74.
- The dataset serves as a valuable resource for training and evaluating NER models within GDPR-compliant privacy policies.

8. FUTURE WORK

- Augmenting the dataset by manually annotating additional privacy policies from various online platforms.
- Utilizing semi-supervised learning techniques to expand the dataset with less manual effort.
- Integrating Relationship Extraction (RE) to capture complex relationships between entities in privacy policies.
- Enhancing the dataset with RE annotations to provide a more comprehensive understanding of privacy policies.

UNIVERSITÄT PASSAU

THANK YOU