On GDPR Compliance of Companies' Privacy Policies

Nicolas M. Müller(✉), Daniel Kowatsch, Pascal Debus, Donika Mirdita, and Konstantin Böttinger

Fraunhofer AISEC, Parkring 4, 85748 Garching near Munich, Germany
{nicolas.mueller,daniel.kowatsch,pascal.debus, donika.mirdita,konstantin.boettinger}@aisec.fraunhofer.de

Abstract. We introduce a data set of privacy policies containing more than 18,300 sentence snippets, labeled in accordance to five General Data Protection Regulation (GDPR) privacy policy core requirements. We hope that this data set will enable practitioners to analyze and detect policy compliance with the GDPR legislation in various documents. In order to evaluate our data set, we apply a number of NLP and other classification algorithms and achieve an F_1 score between 0.52 and 0.71 across the five requirements. We apply our trained models to over 1200 real privacy policies which we crawled from companies' websites, and find that over 76% do not contain all of the requirements, thus potentially not fully complying with GDPR.

Keywords: GDPR data set · GDPR compliance · Natural language processing

1 Introduction

In May 2018, the European Union implemented the General Data Protection Regulation, a new data and privacy protection piece of legislation that applies to everyone within the EU and EEA. This new legislation aims to unify the definition of data protection and privacy related to individuals and forces all the companies that operate in the EU and EEA to provide all the rights, disclaimers and precautions, as defined in the legislation, to every individual in Europe that uses their services and platforms. As a result, companies and businesses from all over the world, large and small, need to update their privacy policies to reflect the GDPR requirements, or shut down their operations in the EU/EEA if they disagree with the legislation. Failure to comply can result in hefty fines.

The GDPR was written and passed in 2016 but it was officially implemented only in May 2018 in order to give businesses a grace period to familiarize themselves with the legislation and make the appropriate changes and adaptations in their own policies. However, not every business entity has made the appropriate changes and a non-negligible fraction of those who have tried to adapt, fall short

K. Ekštein (Ed.): TSD 2019, LNAI 11697, pp. 151–159, 2019.
https://doi.org/10.1007/978-3-030-27947-9_13

on reflecting the necessary policy changes and are thus open to potential fines and/or lawsuits [4,11].

In this paper, we address this situation on several levels, and make the following contribution.

- We present a new labeled data set[1] containing over 18,300 sentence snippets, each labeled with respect to its compliance with five GDPR core privacy policy requirements.
- We show the validity of our data set by designing classifiers which achieve 0.52–0.71 F_1 score between the five classification tasks.
- We evaluate the state of compliance with GDPR 'in the wild': We check over 1200 privacy policies crawled from real companies' websites against our five requirements, and find that at least 76% do not comply with at least one of our core requirements.

Section 2 briefly describes related work. In Sect. 3, we introduce the policy data set in detail and explain its format, class labels and characteristics. Section 4 will introduce the methods we use to generate and rate our policy embeddings and in Sect. 5, we describe our experiments and results. In Sect. 6, we analyze the goodness of our data set and present our conclusions as well as our ideas for further research on this topic.

2 Related Work

There is a considerable amount of research in the area of legal text processing using NLP techniques. However, most of the work is focused on methodologies for doing legal text analysis, automatic rule extraction and summarization [5,7,10]. Other projects have looked at creating annotating tools for extracting complex document rules, regulations, rights and obligations. One of these tools is "Gaius-T", whose performance is satisfactory but it does not provide statistically relevant improvements compared to a human annotator [9]. To the best of our knowledge there is no prior work that analyzes legal compliance on massive user generated policies, and no other work that uses the GDPR legislation as legal text of interest.

3 Privacy Policy Data Set

In this section we introduce our data set. We provide a detailed description and an overview of its characteristics. We also describe the generation and labeling process.

[1] Obtainable at: http://git.aisec.fraunhofer.de/projects/GDPRCOM/repos/on-gdpr-compliance.

3.1 Data Set Generation and Labeling

We obtain our seed data by automatically crawling and storing over 1200 policies in the English language. This set of policies is used for a twofold purpose. First, we use it to train word embeddings (for motivation and details, see Sect. 4). Second, we create our labeled data set for policy compliance rating by manually analyzing and labeling a subset of these policies. This manually labeled set comprises 250 individual policies, containing over 18,300 natural sentences. For legal reasons, we have anonymized the data set, e.g. we have scrambled all numbers and substituted names, email addresses, companies and URLs with generic replacements (e.g. 'company_42645').

We measure policy compliance using five handpicked policy requirements as described in Table 1. Every sentence is assigned a binary score for each class: 0 if the sentence contains no information related to the label, 1 if the sentence discloses class-relevant information. We choose the five requirements in Table 1 because we feel that they represent core requirements of GDPR: They are generic and easily identifiable, which is why we feel that they provide a good overview of how GDPR-compliant a given policy is.

Table 1. The five GDPR requirements we chose to evaluate privacy policy compliance.

No.	Category	Required content in privacy policy	Source
1	DPO	Contact details for the data protection officer or equivalent	[6] 2b/a
2	Purpose	Disclosure of the purpose for which personal data is or is not used for	[6] 2b/b
3	Acquired data	Disclosure that personal data is or is not collected, and/or which data is collected	[6] 2a
4	Data sharing	Disclosure if 3rd parties can or cannot access a user's personal data	[6] 2b/c, d
5	Rights	Disclosure of the user's right to rectify or erase personal data	[6] 2b/f

We use the following set of guidelines.

1. For the **DPO** class, we define a sentence as compliant (assigned the class 1) if the Data Protection Officer or an equivalent authority is named, or contact details of a similar authority are provided.
2. The requirement **Purpose** is considered fulfilled if purpose for processing is stated. Generic purposes such as 'we may use your personal data for any purpose allowed by the law' do not count.
3. The requirement **Acquired data** is considered fulfilled if the sentence informs on the data collected (phone number, first and last name, address, ...), but also if the sentence informs the reader *that* personal data is collected (for more, see [6] 2a).

4. When analyzing **Data sharing**, we label sentences positive that state either of the following: Personal data is shared (a) with other companies, (b) is shared (or not shared) with the public, (c) is transferred to a third country.
5. When labeling for **Rights**, we narrow our threshold down to two GDPR specific definitions: the rights of a user to have his information rectified or deleted. The disclosure of other rights such as transferability is not counted here!

Note the following caveats: First, for classes **Purpose**, **Acquired Data** and **Data Sharing**, the data in question has to be explicitly marked as personal data (e.g. ‘your information’, ‘your data’). Anonymized information and cookies are not considered personal data.

Second, for all requirements except **DPO**, we label sentences positive if they provide any information related to the class in question, which also includes ‘negative’ information. For example, if a sentence clearly states that data would *not* be shared with some entity or would not be used for specific purposes, we consider this a proper disclosure and assign a positive label.

Third, note that we label sentences as positive if they refer to some list or enumeration of information relevant to the class in question. For example, the sentence ‘We use you data for the following purposes:’ is compliant, even though the sentence itself does not contain the purposes.

3.2 Data Set Statistics

In this section, we provide detailed information on the labeled data set. Table 2 shows the number of sentences with a positive label for each of the GDPR requirements, as well as the total number of sentences and documents. Duplicates have already been removed, even if they originate from different documents. Considerable class imbalance can be observed, which is due to the nature of privacy policies. Section 3.3 will detail this issue.

Table 3 provides insight into the GDPR compliance of the set of policies that make up our labeled data set. About 37% of our policies are fully compliant over our five core requirements, 27% comply to four out of five requirements and 1.6% of our policies do not fulfill any of the five core requirements. This analysis foreshadows the results of Sect. 5, where we generalize to all of our 1200 crawled policies and find that a significant percentage does not cover all five of our core requirements.

Table 4 provides more information regarding the occurrence of the different classes in our labeled data set. The *Coverage* column gives the percentage of documents where the individual classes occur. The *Sentences per Doc* column shows the average number of sentences per class per document.

3.3 Class Imbalance

As we can see from Table 2, the classes are considerably imbalanced. For example, for every sentence compliant with the class **DPO**, the data set contains 50 non-compliant sentences. However, this imbalance is to be expected. Privacy policies

Table 2. Data point counts.

No. of labeled documents	250
No. of sentences	18397
DPO	363
Purpose	971
Acquired data	558
Data sharing	904
Rights	299

Table 3. Compliance ratios.

Compliance	Ratio
No class	1.6%
One class	5.2%
Two classes	10.0%
Three classes	19.6%
Four classes	26.8%
Full compliance	36.8%

Table 4. Average frequencies.

Label	Coverage	Sentences per Doc
DPO	63.2%	1.54
Purpose	88.8%	4.52
Acquired data	77.6%	2.57
Data sharing	84.8%	4.19
Rights	60.8%	1.50

are supposed to contain a multitude of information, which is why individual requirements will be represented by a few sentences only.

High class imbalance can be a problem in machine learning [8], which is why we take the following measures. First, instead of accuracy, we use F_1 score as evaluation metric for model selection and evaluation. Second, we use class weights during training, where classes that have a smaller representation get a higher weight in order to even out the representation. Third, we use data up-sampling.

4 Methods

In this section, we introduce the algorithms we use to learn from our data set. Our learning pipeline is as follows. First, we generate sentence embeddings and map the privacy policies under test to numerical vector representation. Second, we use supervised learning algorithms to classify a given sentence as either compliant to a given class, or not. We train and evaluate our classifiers for each class individually, resulting in five independent classifiers.

4.1 Sentence Embeddings

In this section, we detail how we map textual data to vector representation. To this end, we use Word2Vec [12], FastText [1], and ELMo [14], three highly popular word embedding techniques. We experiment with both pre-trained models as well as with training our own embeddings, using 1.200 privacy policies crawled of the web as training data. However, we find pre-trained models to yield superior performance compared to the models we train ourselves. Thus, we exclusively use pre-trained embeddings, whose results we report in the following sections. Specifically, we use the Google News Negative 300 Slim embeddings[2], Facebook's official FastText weights[3] and ELMo embeddings from Tensorflow Hub[4].

4.2 Classification Models

In this section, we introduce the classifiers we use for rating policy compliance. We evaluate Support Vector Machines, Logistic regression and Neural Networks. In order to counter the heavy class imbalance, we use SMOTE up-sampling [2] in conjunction with neural networks. For Logistic regression and SVM, we use balanced class weights to counter the class imbalance.

We train all of the classification models using 3-fold cross-validated grid search to find the optimal set of hyperparameters. Each class is trained and evaluated independently. As evaluation metric, we use the F_1 score, which is the harmonic mean of precision and recall. We use *sklearn*'s implementation [13] of SVM and Logistic Regression, and *keras* [3] to implement the Neural Networks.

5 Experiments

In this section we report the performance of our supervised classifiers on our labeled data set. We then apply the best performing classifiers on a large set of previously unseen privacy policies and evaluate GDPR compliance.

5.1 Classifier Evaluation

Table 5 provides a performance breakdown over all classes, algorithms, and embedding models. We report the test F_1 score of the classifiers with the hyper parameters found via cross-validated grid-search on the train set. We highlight the table cells which show the best classifier per class. Based these results, we find a clear dominance of ELMo embeddings over FastText over Word2Vec. We observe that for most classes, SVM and Neural Networks yield the same performance.

The best hyper parameters for SVM always include the following *sklearn* parameters: $gamma$: $scale$ and $class\_weights$: $balanced$, and either rbf or $poly$ kernel. The best performing neural networks contain two hidden layers with 40 Neurons each, 10% dropout and $relu$ activations.

[2] https://github.com/eyaler/word2vec-slim.

[3] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.en.300.bin.gz.

[4] https://tfhub.dev/google/elmo/2.

Table 5. F_1 test score for all models and embeddings, using the best hyper parameters found via grid search.

	ElMo			FT			W2V		
	LR	NN	SVM	LR	NN	SVM	LR	NN	SVM
Acquired data	0.47	0.52	0.51	0.31	0.49	0.46	0.27	0.48	0.48
DPO	0.59	0.61	0.64	0.38	0.60	0.60	0.37	0.51	0.60
Data sharing	0.55	0.63	0.63	0.44	0.58	0.58	0.42	0.57	0.58
Purpose	0.46	0.54	0.54	0.32	0.46	0.48	0.31	0.48	0.46
Rights	0.61	0.67	0.67	0.31	0.57	0.57	0.35	0.65	0.71

5.2 Experimental Analysis of Unseen Privacy Policies

In this section, we present and interpret the results we obtain when classifying unseen privacy policies from real companies' websites for GDPR-compliance according to our five requirements. Our test data set contains 1,200 documents. For each document, we estimate whether each of our requirements is fulfilled. To this end, we use ELMo embeddings and SVM classifiers, which provided the best or close to best results according to Table 5.

Table 6. Coverage for unseen policies.

Class	Coverage
DPO	68.07%
Purpose	74.72%
Acquired data	62.01%
Data sharing	76.39%
Rights	39.48%

Table 7. Compliance ratios.

Compliance	Ratio
No class	9.56%
One class	8.98%
Two classes	9.81%
Three classes	18.04%
Four classes	30.09%
Full compliance	23.25%

Table 6 provides an estimate of how many of the privacy policies fulfill each of the classes. Our models appear to have good generalization capabilities. A comparison to Table 4 shows a lack of overfitting and overall proportionate results. We observe a difference in coverage values between less than 5% for class **DPO** and about 20% for class **Rights**.

Table 7 shows the compliance rate of our test data set of 1,200 unseen policies. We can see that about 76% are not fully compliant with our requirements, whereas 9.5% do not cover any requirement at all.

6 Conclusion

In this paper, we introduced a novel labeled data set of privacy policies for the purpose of studying GDPR compliance. This data set contains 250 privacy policies and a total of over 18,300 sentences labeled over the five classes **DPO**, **Purpose**, **Acquired Data**, **Data Sharing**, **Rights** as described in Table 1. We apply a comprehensive set of NLP algorithms in combination with supervised learning to analyze the soundness of our data set and build a framework that can rate GDPR compliance of privacy policies. We achieve an F_1 score of 0.52–0.71 between the five classes, indicating that while there is room for improvement with respect to the classification algorithms, our data set may be useful in real-word tasks.

7 Future Work

Future work in this project includes the growth of our labeled data set. Adding more data should boost the performance of our classifiers, and allow for the use of more advanced networks such as RNNs. In addition to this, other classifiers such as Label Propagation could be evaluated. Finally, we would also like to expand our set of legal requirements for compliance, which would make our compliance rating more comprehensive and expressive.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
2. Bowyer, K.W., Chawla, N.V., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. CoRR abs/1106.1813 (2011). http://arxiv.org/abs/1106.1813
3. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
4. Deloitte: Deloitte general data protection regulation benchmarking survey (2018). https://www2.deloitte.com/content/dam/Deloitte/be/Documents/risk/emea-gdpr-benchmarking-survey.pdf
5. Dragoni, M., Villata, S., Rizzi, W., Governatori, G.: Combining NLP approaches for rule extraction from legal documents. In: 1st Workshop on MIning and REasoning with Legal Texts (MIREL 2016) (2016)
6. Gowling WLG: Checklist for tasks needed in order to comply with GDPR. https://gowlingwlg.com/GowlingWLG/media/UK/pdf/170630-gdpr-checklist-for-compliance.pdf. Accessed 31 Mar 2019
7. Hachey, B., Grover, C.: Extractive summarisation of legal texts. Artif. Intell. Law **14**(4), 305–345 (2006)
8. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**, 429–449 (2002)
9. Kiyavitskaya, N., et al.: Automating the extraction of rights and obligations for regulatory compliance. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) ER 2008. LNCS, vol. 5231, pp. 154–168. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87877-3_13
10. Lame, G.: Using NLP techniques to identify legal ontology components: concepts and relations. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 169–184. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-32253-5_11

11. Law.Com: Over half of companies are far from GDPR compliance, report finds (2018). https://www.law.com/corpcounsel/2018/10/19/over-half-of-companies-are-far-from-gdpr-compliance-report-finds/
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
13. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
14. Peters, M.E., et al.: Deep contextualized word representations. CoRR abs/1802.05365 (2018). http://arxiv.org/abs/1802.05365