




# Privacy Policy Classification with XLNet (Short Paper)

Majd Mustapha<sup>(✉)</sup> , Katsiaryna Krasnashchok , Anas Al Bassit ,  
and Sabri Skhiri 

EURA NOVA, 1435 Mont-Saint-Guibert, Belgium  
{majd.mustapha,katherine.krasnoschok,anas.albassit,  
sabri.skhiri}@euranova.eu

**Abstract.** Popularization of privacy policies has become an attractive subject of research in recent years, notably after General Data Protection Regulation came into force in the European Union. While GDPR gives Data Subjects more rights and control over the use of their personal data, length and complexity of privacy policies can still prevent them from exercising those rights. An accepted way to improve the interpretability of privacy policies is through assigning understandable categories to every paragraph or segment in said documents. Current state of the art in privacy policy analysis has established a baseline in multi-label classification on the dataset containing 115 privacy policies, using BERT Transformers. In this paper, we propose a new classification model based on the XLNet. Trained on the same dataset, our model improves the baseline F1 macro and micro averages by 1–3% for both majority vote and union-based gold standards. Moreover, the results reported by our XLNet-based model have been achieved without fine-tuning on domain-specific data, which reduces the training time and complexity, compared to the BERT-based model. To make our method reproducible, we report our hyper-parameters and provide access to all used resources, including code. This work may therefore be considered as a first step to establishing a new baseline for privacy policy classification.

**Keywords:** Privacy policy · Multi-label classification · Deep learning

## 1 Introduction

Despite the rising importance of how personal data is managed and protected, people still routinely skip privacy policy contracts, due to their complexity and length. A simple word count on privacy policies of the biggest digital companies shows that after GDPR came into force in 2018, the length of privacy policy contracts has increased by over 25% on average, peaking at 94% for Wikipedia [13]. With the increase of number of digital services we use, it became less and less enticing to try to understand what is seemingly an endless block of text.

The recent research works have made a considerable progress in helping scientists and end users make sense of the conditions described in privacy policies,

by classifying their segments into understandable pre-defined categories, that users can refer to and compare between policies. These efforts resulted in the creation of several datasets of various detail level, containing diverse categories describing the policies from different aspects, depending on the objective.

This work is developed within the ASGARD project<sup>1</sup>, in particular, its RUNE track, whose objective is supporting the automation of privacy by design. One of the primary tasks of the track is translation of privacy policies, data processing agreements and contracts into a machine-readable format. Such task requires a trustworthy dataset for extraction of policy attributes. We experiment with the OPP-115 dataset [16], which is an accepted gold-standard containing 115 annotated privacy policies. To the best of our knowledge, this is the most detailed and widely used privacy policy dataset in the research community, despite being somewhat outdated and incomplete for usage in the GDPR-specific context [6].

In this paper, we present a privacy policy classification model based on XLNet [17] and showcase its performance in comparison to the baseline, established by Najmeh Mousavi Nejad et al.<sup>2</sup> [10] using BERT on OPP-115 dataset with two gold standards: majority vote and union-based. Our goal is to strengthen the baseline results using the latest advancements in Deep Learning and Natural Language Processing, as well as to demonstrate the performance of pre-trained XLNet in legal domain. Our approach of applying XLNet for privacy policy classification outperforms the state of the art in terms of macro/micro average F1-scores by 2%/1% for the majority vote and 3%/3% for the union-based gold standard. This result has been achieved without fine-tuning our XLNet-based model on domain-specific data, comparing to the fine-tuned BERT-based model in [10]. We make sure to guarantee reproducibility of our results through keeping the same splits as the baseline [10] and sharing the hyperparameters and code<sup>3</sup>. This work may therefore be considered as a first step to establishing a new baseline for privacy policy classification with OPP-115 dataset.

The paper is structured as follows: in Sect. 2 we lay out the research efforts in privacy policy analysis; in Sect. 3 we describe the model and how it differs from the BERT-based baseline model; Sect. 4 reports our results in privacy policy classification; we discuss our findings in Sect. 5; finally, Sect. 6 concludes the paper and outlines our plans for the future work.

## 2 Related Work

After GDPR has been enforced EU-wide, interest towards privacy policy analysis has increased significantly, which is evident by the great number of privacy and GDPR-related research projects in the EU and worldwide. Among the most prominent of them, the Usable Privacy Policy Project<sup>4</sup>, started long before the

<sup>1</sup> Supported and funded by the Walloon region, Belgium.

<sup>2</sup> To be published in the proceedings of The 35th International Conference on ICT Systems Security and Privacy Protection (2020).

<sup>3</sup> <https://github.com/euranova/privacy-policy-classification-xlnet>.

<sup>4</sup> <https://usableprivacy.org/>.

GDPR, aims to benefit users through demystifying privacy policies. OPP-115 dataset [16] has been created in the context of the project, and became the first of its kind, with fine-grained annotations on paragraph level. Several other useful datasets have also been released for the same project [12, 18]. In our work we make use of OPP-115, as it is the most used dataset in the privacy policy research, due to its detail level and rigorous annotation procedure.

Another outstanding project in the field of improving interpretability of privacy policies is Polisis [7] – a framework that categorizes, visualizes, and explains the contents of a policy to an end user in an interactive manner. The authors have trained their classification model on OPP-115, and reported their results. In this paper we do not compare our model to Polisis’ CNN-based model, since the current BERT-based state of the art already outperforms it.

Beyond the research community, we can note “Terms of Service; Didn’t Read” (ToS;DR)<sup>5</sup> project, which utilizes crowd-sourcing efforts to evaluate and classify terms of service and privacy policy documents in the context of their fairness to the users and how much concern they raise for data privacy and security. The ratings given to various services and websites help end users grasp the overall meaning and important notions in the policies, though the categories are less detailed than the ones the OPP-115 dataset presents.

When it comes to classification of textual data, until very recently, state of the art relied mostly on the variations of Recurrent Neural Networks [3, 8]. However, the inherent sequential nature of recurrent models is what limits their ability to process long sentences and stands in the way of faster parallel training. Attention mechanism [2] confronted the problem by modeling dependencies regardless of the distance between the sequence elements. Consequently, Transformers [14] were designed to speed up the training for neural machine translation, through reducing sequential computation with multiple self-attention heads. The Bidirectional Encoder Representations from Transformers (BERT) [5] improved upon the limitations of existing work in pre-trained contextual representations [9, 11] by using deeply bidirectional contextualization. BERT was the first generic representation model that achieved state-of-the-art performance on a large array of sentence-level and token-level tasks, outperforming many task-specific models [5]. In the context of our work, the latest and best reported performance on the OPP-115 dataset until now has been achieved by Najmeh Mousavi Nejad et al. [10] with a model based on fine-tuned BERT, which we adopt as a baseline.

### 3 XLNet Privacy Policy Classification Model

In order to compare fairly to the state of the art, we use the OPP-115 dataset with the same splits as in [10], on which we train our classifier, consisting of a pre-trained XLNet<sup>6,7</sup> and a dense layer for classification. In this Section, we lay

<sup>5</sup> <https://tosdr.org/about.html>.

<sup>6</sup> <https://github.com/huggingface/transformers>.

<sup>7</sup> <https://github.com/kaushaltrivedi/fast-bert>.

out the background and justify the decisions made for our classification model, by discussing the differences between XLNet and BERT.

### 3.1 Transformer-XL and XLNet

A limitation of vanilla Transformers is in stateless computations that put an upper limit on the distance of relationships they can model [1]. The Transformer-XL [4] is an extension of the Transformer that overcomes this shortcoming by caching the hidden states of the previous sequence and passing them as keys/values when processing the current sequence. It also introduces relative positional embeddings that encode relative distances between words and allow the model to compute the attention score for words that are before and after the current word.

XLNet [17] changed the way a language modeling problem is approached. It is an auto-regressive language model that outputs the joint probability of a sequence of tokens with recurrence. It calculates the probability of a word, conditioned on all possible permutations of words in a sentence, as opposed to just those to the left or the right of the target word. The model achieves state-of-the-art performance on the GLUE benchmark [15], trained on a large corpus.

### 3.2 XLNet vs BERT

Despite its strong performance across the multitude of tasks, BERT has attracted criticism due to the following flaws [17]:

- In the Transformer architecture the model can acquire context information exclusively within the boundaries of the maximum input sequence length, so a longer document would be divided into independently processed segments.
- BERT suffers a discrepancy between fine-tuning and pre-training, when it comes to predicting masked tokens: during pre-training, tokens are replaced with the [MASK] symbol, though, it never appears in downstream tasks.
- When predicting masked tokens, BERT disregards the dependencies between them, thus reducing the number of dependencies it can learn at once.

The sequence length constraint is tackled by XLNet due to the features of Transformer-XL, whose Recurrence Mechanism and Relative Positional Encoding help capture long-term dependencies for longer documents. The model caches the hidden state sequence, computed from the previous segment, and reuses it as an extended context, when processing the next segment. This additional input allows the network to exploit historical information, and still keep the gradient within a segment. While BERT encodes context positions statically, Relative Positional Encoding of Transformer-XL allows for the encoding of positions in a relative distance from the current token at each attention module. The aim is to accommodate the Recurrence Mechanism and avoid having tokens from different positions with the same positional encoding.

Transformer-XL only holds unidirectional context, predicting current token based on sequential context on its left or its right. However, it solves the issue by

introducing the Permutation Language Modeling objective: instead of predicting tokens in sequential order, it follows a random permutation order. Only the last tokens in a factorization order are chosen for training to reduce optimization difficulty that comes from working with permutations.

Building on the information above, we believe that applying XLNet to the downstream task of privacy policy classification holds the potential of improving the current baseline results achieved with the BERT-based model in [10].

## 4 Evaluation

To evaluate our approach, we follow Najmeh Mousavi Nejad et al. [10] and train our XLNet-based classifier on the Online Privacy Policies (OPP-115) dataset. A comprehensive description of the dataset and its categories can be found in [16], and the gold standards with their label distributions are presented in [10]. Thus, here we briefly mention the key aspects of the dataset that are necessary for the interpretation of the results.

OPP-115 consists of 115 privacy policies, manually annotated on a paragraph level, resulting in 3 792 paragraphs, 10 high-level classes and 22 distinct attributes. Like the majority of previous works, we are only considering the high-level categories for classification, 12 exactly<sup>8</sup>. Therefore, we have a 12-class multi-label classification task at hand. In order to establish a firm comparison to the state-of-the-art results, we apply the same splits used by Najmeh Mousavi Nejad et al. [10]: the authors reported that they randomly partitioned splits, according to Machine Learning best practices, into a ratio of 3:1:1 for train, validation and test, respectively. For the same purpose, we also evaluate on the two gold standards, considered by the baseline model: the majority vote and union-based. We report the resulting F1 values of our XLNet-based model in Table 1, in comparison to BERT-based model performance reported in [10].

Table 1 shows that XLNet improves both baselines - BERT and BERT fine-tuned - without the need of fine-tuning on the domain-specific data. These improvements can be explained by the architectural differences between XLNet and BERT, mentioned in Sect. 3, and additionally, by the fact that XLNet has been trained on a bigger corpus, that includes the training data of BERT. Therefore, it works with bigger vocabulary and moreover, it generalizes better. Another factor that we believe affected the performance for the better, is Transformer-XL's Recurrence Mechanism and Relative Positional Encoding, that help capture long-term dependencies for long documents and sentences. This feature is especially important in analysis of legal documents, such as privacy policies, which tend to have long and complicated sentence and paragraph structure. As evident from Table 1, in total, our XLNet-based model outperforms the state of the art by 2%/1% for the majority vote gold standard and 3%/3% for the union-based gold standard, for macro/micro F1 average scores, respectively, while keeping the tendency for micro- to outperform macro-averages, mentioned also in [10].

<sup>8</sup> We follow the baseline [10], where the *Other* category was broken down into its 3 underlying attributes.

**Table 1.** F1 values in % for 2 baseline models from [10] models and our model (in bold) on the two gold standards with a threshold = 0.5 (V - validation; T - test)

Labels	Majority-vote gold standard						Union-based gold standard					
	BERT		BERT FT		<b>XLNET</b>		BERT		BERT FT		<b>XLNET</b>	
	V	T	V	T	V	T	V	T	V	T	V	T
First Party Collection & Use	87	88	<b>88</b>	91	89	90	83	84	87	86	85	87
Third Party Sharing & Collection	86	85	87	90	89	88	79	82	83	86	83	89
User Access, Edit and Deletion	82	63	77	73	81	76	54	49	56	65	70	73
Data Retention	40	33	54	56	62	64	36	68	62	71	75	73
Data Security	87	82	54	56	89	81	71	80	73	76	76	78
International/Specific Audiences	94	81	87	80	95	84	87	93	92	92	90	90
Do Not Track	80	100	95	83	80	100	80	60	100	92	80	93
Policy Change	80	88	80	100	89	89	75	78	77	80	70	84
User Choice/Control	75	81	85	90	81	77	64	63	66	65	64	69
Introductory/Generic	75	76	78	79	82	81	74	68	73	67	67	74
Practice Not Covered	18	32	35	35	56	42	44	46	45	48	45	56
Privacy Contact Information	79	80	79	78	84	80	75	71	83	78	80	83
<b>Macro Averages</b>	74	74	77	79	81	<b>81</b>	68	70	75	76	77	<b>79</b>
<b>Micro Averages</b>	81	82	83	85	86	<b>86</b>	73	74	77	77	78	<b>80</b>

If we compare to base BERT, the difference is more remarkable, encoding the performance gap between “pure” BERT and XLNet: 7%/5% for the majority vote and 9%/6% for union-based gold standard (macro/micro F1).

## 5 Discussion

Looking at the F1 values per label, we can note that XLNet outperforms both BERT and BERT fine-tuned on most of the categories, with an impressive increase in certain cases. A good example is *Data Retention* class, whose F1 metric improved greatly (from 56% to 64%) for majority-vote gold standard, but not so much for the union-based, where we had more than twice as much training examples<sup>9</sup>. Another category that exhibited poor performance for the BERT-based models is *Practice Not Covered*: as noticed by the authors, this class covers broad range of topics and vocabulary, which makes it harder for the classification model to learn. Interestingly, XLNet improves F1 values on this class significantly in both gold standards, while still demonstrating better performance on union-based label set. From these examples we can conclude that XLNet has the potential of improving classification results for the underrepresented or “tough to learn” classes, even with small amount of examples to learn from.

As we noted before, for this paper we did not consider fine-tuning XLNet (which would take considerable time and resources in preparation and training) as the pre-trained version has already given us the desired improved performance, comparing even to fine-tuned BERT, let alone the base one. However, we have

<sup>9</sup> For the label distribution in the two gold standards see [10].

reasons to believe that, just like with BERT, fine-tuning XLNet on the domain-specific data (a big set of privacy policies) should result in even higher F1 values. Currently, we leave this step for the future work.

Additionally, we would like to point out that the training of our model did not require any significant resources or time, in fact, the training configuration is the same as for the BERT-based model, for the most part. Hence, the improved results have been achieved without a sacrifice in training resources.

## 6 Conclusion and Future Work

In this paper, we demonstrated the performance of recently released XLNet model in legal/privacy domain, where this kind of model has not been applied, to the best of our knowledge. We evaluated an XLNet-based multi-label classification model on the OPP-115 dataset, with the goal of establishing a new baseline for privacy policy analysis. Our experiments with a pre-trained XLNet showed that it outperforms BERT on this particular domain-specific task, and moreover, it does so without the need to be fine-tuned on the domain-specific data.

In terms of the future work, we plan to experiment with fine-tuning XLNet on a large set of privacy policies, and we expect this step to further improve the results. As for the next phases in terms of the ASGAR project, we intend to use the model and the classification results to translate privacy policies into a machine-readable representation, to be used in the downstream applications, such as compliance checking and access control for business requests. For this purpose, the dataset and annotations will need to be enriched with missing concepts, including GDPR-specific attributes, such as various legal basis terms. It becomes increasingly important to be able to extract legal basis, as the majority of new and updated policies mention it for the purpose of being GDPR-compliant, yet the current version of the dataset contains only a subset of the legal bases mentioned in the GDPR. Therefore, our future work will focus on improving both the classification model and the dataset, in order to obtain the high quality representations of policies and contracts.

## References

1. Al-Rfou, R., Choe, D., Constant, N., Guo, M., Jones, L.: Character-level language modeling with deeper self-attention. CoRR abs/1808.04444 (2018)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings (2015)
3. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning (2014)

4. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers, pp. 2978–2988. ACL (2019). <https://doi.org/10.18653/v1/p19-1285>
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. ACL (2019). <https://doi.org/10.18653/v1/n19-1423>
6. Gallé, M., Christofi, A., Elshar, H.: The case for a GDPR-specific annotated dataset of privacy policies. In: AAAI Symposium on Privacy-Enhancing AI and HLT Technologies (2019)
7. Harkous, H., Fawaz, K., Lebrete, R., Schaub, F., Shin, K.G., Aberer, K.: Polis: automated analysis and presentation of privacy policies using deep learning. In: 27th {USENIX} Security Symposium, {USENIX} Security 2018, pp. 531–548 (2018)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–80 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
9. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 1: Long Papers, pp. 328–339. ACL (2018). <https://doi.org/10.18653/v1/P18-1031>
10. Mousavi, N., Jabat, P., Nedelchev, R., Scerri, S., Graux, D.: Establishing a strong baseline for privacy policy classification. In: IFIP International Conference on ICT Systems Security and Privacy Protection (2020)
11. Peters, M., et al.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237. ACL, New Orleans (2018). <https://doi.org/10.18653/v1/N18-1202>
12. Sathyendra, K.M., Wilson, S., Schaub, F., Zimmeck, S., Sadeh, N.M.: Identifying the provision of choices in privacy policy text. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, pp. 2774–2779. ACL (2017)
13. Sobers, R.: The average reading level of a privacy policy (2020). <https://www.varonis.com/blog/gdpr-privacy-policy/>
14. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)
15. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint [arXiv:1804.07461](https://arxiv.org/abs/1804.07461) (2018)
16. Wilson, S., et al.: The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1330–1340 (2016)
17. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pp. 5754–5764 (2019)
18. Zimmeck, S., et al.: MAPS: scaling privacy compliance analysis to a million apps. *PoPETs* **2019**(3), 66–86 (2019). <https://doi.org/10.2478/popets-2019-0037>