Establishing a Strong Baseline for Privacy Policy Classification

Najmeh Mousavi Nejad[1,2], Pablo Jabat[3], Rostislav Nedelchev[1], Simon Scerri[2], and Damien Graux[4]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
nejad@cs.uni-bonn.de, rostislav.nedelchev@uni-bonn.de
http://sda.cs.uni-bonn.de

[2] Fraunhofer Intelligent Analysis and Information Systems (IAIS), Germany
simon.scerri@iais.fraunhofer.de
https://www.iais.fraunhofer.de

[3] Company Watch Ltd.
pjabat@companywatch.net
https://www.companywatch.net/

[4] ADAPT Centre, Trinity College Dublin, Ireland
damien.graux@adaptcentre.ie
https://www.adaptcentre.ie/

Abstract. Digital service users are routinely exposed to Privacy Policy consent forms, through which they enter contractual agreements consenting to the specifics of how their personal data is managed and used. Nevertheless, despite renewed importance following legislation such as the European GDPR, a majority of people still ignore policies due to their length and complexity. To counteract this potentially dangerous reality, in this paper we present three different models that are able to assign pre-defined categories to privacy policy paragraphs, using supervised machine learning. In order to train our neural networks, we exploit a dataset containing 115 privacy policies defined by US companies. An evaluation shows that our approach outperforms state-of-the-art by 5% over comparable and previously-reported F1 values. In addition, our method is completely reproducible since we provide open access to all resources. Given these two contributions, our approach can be considered as a strong baseline for privacy policy classification.

Keywords: Privacy Policy · Multi-label Classification · Deep Learning.

1 Introduction

Various studies indicate that, despite their proliferation, a majority of consumers still skip privacy policy consent forms due to the difficulty required for lay users to comprehend their contents. In fact, a recent study called “The Biggest Lie on the Internet” reported that only around a fourth of participants read privacy policies, and they only invest just over a minute to do so [16]. Moreover, these statistics are probably lower outside of laboratory conditions. Another survey showed that if users were to read the privacy policies of all services they visit on the Internet, they would need on average

244 hours each year which is almost more than half of the average time a user spends on the Internet [13].

To assist end-users with consciously agreeing to the conditions, we consider Natural Language Processing (NLP) and Machine Learning (ML) methods and apply them to classify privacy policy paragraphs into pre-defined categories for easier comprehension. Our efforts seek to build on the results of two earlier dominant studies in the literature. The first is the OPP-115 dataset, which contains 115 privacy policies at paragraph level, each of which includes fine-grained annotations from 3 experts [23]; e.g., the paragraph in Figure 1 from the Amazon policy[1] is annotated with two classes: *User Access, Edit & Deletion* and *Data Retention*. The second study which inspired our research is the effort by *Polisis* to build a Convolutional Neural Network (CNN) model exploiting OPP-115 [6]. Despite the valuable contribution of these earlier studies, they exhibit one major weakness: reproducibility. Due to a lack of information on the exact ML dataset splits used, and the lack of a common gold standard in the literature, subsequent studies have created their own. This makes it difficult to collectively interpret and compare the different results. A major contribution of the efforts presented here is our provision of a strong and completely reproducible baseline for future research.

> **“** [. . .] You can add or update certain information on pages such as those referenced in the *Which Information Can I Access?* section. When you update information, we usually keep a copy of the prior version for our records. [. . .] **”**
>
> – *User Access, Edit and Deletion*
> – *Data Retention*

Fig. 1: Excerpt from Amazon privacy notice

More concretely, our contributions are the following:

- A comprehensive set of experiments based on two different gold standards;
- A presentation of a strong baseline for privacy policy classification using NLP and ML that successfully reproduces state-of-the-art findings (though with our self-created data splits and gold standards) and furthermore improves the results by employing the *BERT* framework [4] for the two gold standards;
- Ensuring the reproducibility of our results by providing all resources utilised to generate our conclusions.

Central to our efforts is a multi-label classification problem with 12 classes, which can be used to predict one or more classes for each paragraph of a given privacy policy, based on a neural network and the OPP-115 dataset. We first compiled two gold standards from OPP-115: one based on majority votes (i.e., two or more experts agree on a label); and the other with the union of all expert annotations. The dataset creators [23] considered the majority-vote-based standard, whereas *Polisis* used the union-based,

[1] To retrieve the exact source used: <https://www.amazon.com/gp/help/customer/display.html?nodeId=468496> (Sub-entry *What Choices Do I Have?*) – last accessed March.2nd.2020

with the rationale that disagreements are a result of the experts' high understanding of legal texts and that therefore, none of their annotations should be deemed incorrect.

In order to establish a strong baseline, we compare three models with both gold standards. The first model is a CNN, whose generation is directly comparable to the earlier *Polisis* efforts. The second and third models are based on the *BERT* transformer, a model that has recently gained a lot of attention as a potential superior alternative. To the best of our knowledge, our efforts are the first attempt to produce a reliable and completely reproducible result on privacy policy classification. The results attained demonstrate consistency and significant improvement over the baseline and indicate good reliability: A 77% micro-average F1 on the union-based gold standard, and a 85% micro-average F1 on the majority-based gold standard.

The rest of the paper is divided as follows: in section 2 we compare our approach to the existing studies on privacy policies. Section 3 provides details of the three models. In sections 4 and 5, an extensive set of experiments is presented and discussed. Finally, section 6 concludes this study and suggests future directions towards privacy policy analysis.

2 Related Work

In light of the, now enforced EU-wide, General Data Protection Regulation (`GDPR`), there has been an increasing interest toward privacy policy analysis. Some studies investigated the essential regulatory model, *notice and choice* [10] in web privacy principles [11,17]. Libert monitored data flows on websites and identified third parties who collect and use personal data [11]. Afterward, over 200,000 websites' privacy policies are scanned to determine whether the parties identified, are explicitly mentioned in the page's privacy policy. Furthermore, privacy policies are additionally analyzed to check whether they respect the "Do Not Track" browser setting[2]. In another study, the authors applied NLP and supervised ML to automatically extract control choice excerpts and opt-out hyperlinks from privacy policy documents [17]. In order to evaluate their work, OPP-115 was used and the results showed that ML is feasible, even with the small number of samples for *'User Choice/Control'* category in OPP-115. In contrast to our problem, these approaches have addressed only a specific feature of privacy policies, whereas our method processes the whole document for the benefit of regular end-users.

A few approaches developed a model with supervised ML to measure completeness of privacy policies [5,3]. The dataset used in training, contains a set of pre-defined categories based on privacy regulations and guidelines. Finally the trained model predicts a category for an unseen paragraph. According to the papers, this structure helps users to examine privacy policies faster and allows them to focus on those categories in which they are interested. However, based on our observation, most of online privacy policies use rich HTML representations and therefore offer a basic level of structural view to the end-users. Moreover, to the best of our knowledge none of the corpora were created with the full support of experts, which is an essential prerequisite in legal text processing.

[2] https://en.wikipedia.org/wiki/Do_Not_Track

A prominent group on privacy policy analysis is *Usable Privacy Policy Project*[3], they provided OPP-115, the first comprehensive dataset with fine-grained annotations on paragraph level [23]. The project aims to extract important information for the benefit of regular and expert end users. To do so, a corpus containing 115 privacy policies from 115 US companies was annotated by 3 experts on paragraph level (10 experts in total and 3 experts per document). The annotations in OPP-115 dataset are in two levels: 10 high level categories and 22 distinct attributes. For instance, the high level category *First Party Collection* has 9 low level attributes, some of which are: *Collection Mode, Information Type, Purpose*. Along with the creation of dataset, the authors built different ML models for prediction of high level categories. The gold standard for evaluating the methods was compiled based on majority votes: if two or more experts agreed on a single category, it was considered in the final gold standard. The best reported micro-average F1 is 66% that was achieved with Support Vector Machine (SVM).

Leveraging OPP-115 and deep learning, *Polisis* extracts segments from privacy policies and presents them to users in a visualized format [6]. According to the paper, the union-based gold standard is used for experiments; 65 privacy policies were considered for training and 50 policies were kept for the test set. The authors claim that a successful multi-label classifier should not only predict the presence of a label, but also its absence[4]. They report only macro-averages and further compute the average of F1 and F1-absence and yield 81% average on the test set. Despite the encouraging work done in *Polisis*, we believe that the paper lacks two fundamental elements: there is no validation set involved in training phase; and there is no information on micro-averages.

It is worth to mention that none of the above studies provided their dataset splits and therefore there is no standardized benchmark for privacy policy classification. As a result, in the following sections, first we show how we successfully reproduce *Polisis* results (though with different data splits) and further present two transformer models that significantly outperform *Polisis*.

3 Approach

In order to establish a firm foundation, we attempt to reproduce the work of [6] with additional improvements. To do that, we conduct experiments using word embeddings and a Convolutional Neural Network (CNN). Furthermore, we evaluate Bidirectional Encoder Representations from Transformers (*BERT*) [4] that has state-of-the-art performance on many other text classification tasks.

3.1 Convolutional Neural Network

Pre-trained Word-Embeddings Traditionally, text classifiers have taken advantage of vector representations like bag of words or term-frequency inverse-document-frequency (TF-IDF). However, it is clear that this method has the disadvantage of not retaining the

[3] https://usableprivacy.org/

[4] They also claim that a model that predicts that all labels are present would have 100% precision and recall, which is obviously wrong.

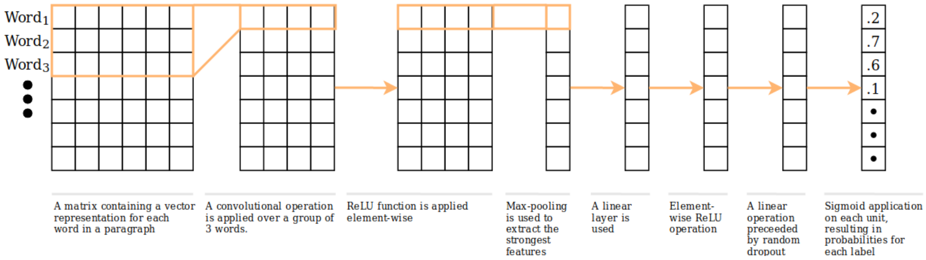
Word$_1$
Word$_2$
Word$_3$
.2
.7
.6
.1
A matrix containing a vector representation for each word in a paragraph
A convolutional operation is applied over a group of 3 words.
ReLU function is applied element-wise
Max-pooling is used to extract the strongest features
A linear layer is used
Element-wise ReLU operation
A linear operation preceeded by random dropout
Sigmoid application on each unit, resulting in probabilities for each label

Fig. 2: CNN Architecture

semantic information depicted by the order of words, as well as the meaning of the single words as independent units and be purely dependent on the context. Thus, we investigate word embeddings.

Word embeddings were initially proposed by [15,2] and were later popularized by [14]. The continuous bag of words (cbow) method, which is a variant of word2vec, creates a numeric representation of words by attempting to predict a given word by considering its neighbors as seen in text. A huge benefit such an algorithm is that, no labeled data is necessary, but only great amounts of correct text.

While word2vec is effective at storing some semantic meaning in a vector representation, it treats words as atomic units and thus, it does not take into consideration the internal structure of words. Such information can be useful for less frequent words or for compound words like rainfall or greenhouse. FastText uses a bag of character n-grams to represent words, where each character n-gram is a vector and all the constituents are summed up to create a representation for the word [7,1].

The aforementioned properties can be useful for the context of privacy policies. Since most openly available word embeddings are trained on news or Wikipedia corpora [1], we utilize fastText to create vector representations that are more suitable for the current task. For that purpose, we used a big corpus of 130k privacy policies scraped from an application store for smart phones. In app stores, applications are required to provide privacy policies. After tokenizing the text with NLTK [19], there are 132 595 084 tokens in total and 173 588 unique ones. We compared the vocabulary between this corpus and two version of OPP-115 that we utilize. We saw that there are 1 072 words which are seen only in OPP-115 majority-vote version, but not in the corpus used for drafting the word vectors. Similarly, for the gold standard containing union of all classes, there were 1 119 out-of-vocabulary (OOV) words. The difference in the amount of OOVs is due to the fact that the majority vote dataset has less paragraphs (when there was no agreement on a single category) and thus, it is less likely that there are unseen words. More details regarding the size of the dataset versions are provided in Section 4. After manual inspection, we concluded that most of the out-of-vocabulary words are names of brands, products, services or their web-addresses. These are completely omitted, since from an intuitive perspective they should not be decisive for the correct detection of a policy class. Hence, the vocabulary is sufficient for the task.

Architecture To tackle the multi-label classification problem, we follow the work of [6] by using a CNN (displayed in Figure 2). The previously explained word embeddings are provided as input to the neural network. A convolutional operation is applied with a context window of 3 words, whose output then passes through a Rectified Linear Activation (ReLU) function. Then, from each context output, only the strongest features are selected by a max-pooling layer, resulting in a single vector that contains the most informative properties of each context, thus the neural network is forced to focus only on certain features that are specific to the current goal. Furthermore, a linear layer followed by a ReLU are applied to create a higher level representation of the collected information. Finally, a linear layer with as many nodes as classes is applied to provide an output in the target dimensions and passed through a sigmoid function to obtain per label probability scores.

The proposed architecture shares a strong resemblance with the work of [8], where a CNN is used for multi-class classification of sentences. However, it lacks a random dropout just before the last linear layer. We conduct experiments with 50% dropout. Additionally, we used Adam [9] optimization algorithm combined with early stopping. The convolutional neural network is optimized using binary cross entropy loss:

$$\ell(x,y) = L = \{l_1,\dots,l_N\}^\top \tag{1}$$

$$l_n = -w_n\left[y_n \cdot \log x_n + (1-y_n)\cdot \log(1-x_n)\right] \tag{2}$$

where $l_1,\dots,l_N$ specify the 12 loss values for each of the 12 possible labels that we have in the dataset. It is being calculated for each, since this is a multi-label classification and we could have any combinations of those. After we have the 12 losses, we take the mean of those 12 to get one scalar number. Furthermore, x is the model prediction, y is the true label, w is the class specific weight which in our case are all 1. For instance, if we consider that our current model assigns probability p to observation o for the *Data Retention* label, the loss function for this specific label will be:

$$loss(DataRetention) = y \cdot \log p + (1-y)\cdot \log(1-p) \tag{3}$$

where y is 1 if observation o is labeled with *Data Retention* in the gold standard and 0 if not.

3.2 Bidirectional Encoder Representations from Transformers

The *BERT* framework [4] uses several layers of transformer encoders [21] to create a bidirectional representation of the tokens in the sequence. The approach operates in two stages: first, the model is pre-trained on large amounts of unlabelled data; second, it is fine-tuned on specific labeled data to solve a downstream problem, which in our case is multi-label classification.

To handle various domains and tasks, *BERT* is using WordPiece [24] tokenization. It provides a reasonable balance between character and subword level information. For example, a model using it, can detect similar suffixes or roots among words. This way, the vocabulary stays within a reasonable size, without having too many entries. The chosen vocabulary size is 30 000 [4].

BERT is pre-trained using two unsupervised tasks. The first one is masked language modeling (MLM), i.e., the model is being taught to predict 15% of the randomly "masked" tokens in a sentence. The masking uses one of three randomly chosen possible ways: 1) in 80% of the cases, a token is replaced with [*MASK*]; 2) in 10% with another random word; and 3) in the remaining 10% no replacement is done [4]. The other unsupervised language modeling task is next sentence prediction (NSP). Every input sequence to the framework always starts with the classification token [*CLS*], which provides a fixed-length representation for the whole input. For NSP, two subsequent sentences from the corpora are concatenated with another separator token, [*SEP*], so that the model is aware of the separation between the two. In 50% of the cases, the second sentence is replaced by another one. Thus, *BERT* is trained to recognize when a pair of sentences appear together in the corpora (or they don't), using the classification token [4].

We use a pre-trained version of $BERT_{BASE}$[5, 6] which has 12 encoder layers, a hidden state size of 768, and 12 attention heads, totaling in 110M parameters. Additionally, we also prepare another fine-tuned version of the language model with our 130K privacy policy corpus[7]. Ninety percent of those were used for training while the remaining ten for validation. We fine-tune the model for three epochs and achieved a cross-entropy loss on the mask languaged model task of 0.1151 and perplexity, 1.1220. Finally, both versions of the approach are trained on the privacy policy classification task and evaluated. For more detail on *BERT*, we would forward the reader to the relevant references [4,21].

4 Evaluation

In pursuance of providing a reliable baseline for privacy policy classification, two gold standards were compiled out of OPP-115 dataset. OPP-115 high-level annotations are divided into 10 classes:

1. *First Party Collection/Use*: how and why the information is collected.
2. *Third Party Sharing/Collection*: how the information may be used or collected by third parties.
3. *User Choice/Control*: choices and controls available to to users.
4. *User Access/Edit/Deletion*: if users can modify their information and how.
5. *Data Retention*: how long the information is stored.
6. *Data Security*: how is users' data secured.
7. *Policy Change*: if the service provider will change their policy and how the users are informed.
8. *Do Not Track*: if and how Do Not Track signals is honored.
9. *International/Specific Audiences*: practices that target a specific group of users (e.g., children, Europeans, etc.)
10. *Other*: additional practices not covered by the other categories.

[5] https://github.com/huggingface/transformers

[6] https://github.com/kaushaltrivedi/fast-bert

[7] The `BertLMDataBunch` class contains `from_raw_corpus` method that takes a list of raw texts and creates `DataBunch` for the language model learner.

Labels	Union						Majority Votes					
	Tr	V	T	Tr(%)	V(%)	T(%)	Tr	V	T	Tr(%)	V(%)	T(%)
First Party Collection & Use	988	243	288	40.8	40.1	38	781	176	250	34.2	30.9	35
Third Party Sharing & Collection	755	204	227	31.1	33.7	30	584	158	203	25.5	27.7	28.4
User Access, Edit and Deletion	155	29	46	6.4	4.8	6.1	101	24	24	4.4	4.2	3.4
Data Retention	111	21	24	4.6	3.5	3.2	50	14	14	2.2	2.4	2
Data Security	251	65	59	10.3	10.7	7.8	139	31	40	6.1	5.4	5.6
International/Specific Audiences	225	67	61	9.3	11.1	8.1	204	41	56	9	7.2	7.8
Do Not Track	22	3	7	1	0.5	0.9	22	6	3	1	1	0.4
Policy Change	118	27	47	4.9	4.4	6.2	73	25	21	3.2	4.4	3
User Choice/Control	405	97	130	16.7	16	17.2	233	48	77	10.2	8.4	10.8
Introductory/Generic	514	137	162	21.2	22.6	21.4	240	72	78	10.5	12.6	11
Practice Not Covered	402	102	138	16.6	16.8	18.2	83	21	25	3.6	3.7	3.5
Privacy Contact Information	207	44	72	8.5	7.3	9.5	129	32	42	5.6	5.6	5.9

Table 1: Label distribution in the two gold standards; Tr:Train; V:Validation; T:Test

Ten experts were hired to create fine-grained annotations and each privacy policy was randomly assigned to 3 of them. OPP-115 comprises 3 792 paragraphs, 10 high-level classes and 22 distinct attributes[8]. Each paragraph was labeled with one or more classes (out of 10). According to the dataset creators, the best agreement was achieved on *Do Not Track* class with Fleiss' Kappa equal to 91%, whereas the most controversial class was *Other*, with only 49% of agreement [23]. The latter category was further decomposed into its attributes: *Introductory/Generic*, *Privacy Contact Information* and *Practice Not Covered*. Therefore, we face a multi-label classification problem with 12 classes. It should be clarified here that computing Fleiss' kappa considering all categories together is not feasible for OPP-115, as annotators differ per policy. Aforementioned, there were 10 experts and each policy was randomly assigned to 3 of them. If 3 experts were the same experts for the whole dataset, it was rational to compute an overall Fleiss's kappa for all 10 categories and between 3 annotators. For this reason, [23] reported Fleiss' kappa per category.

To evaluate our three models, we compiled two gold standards: union-based, which contains all expert annotations; and the majority-vote-based gold standard, where only annotations with an agreement between at least 2 experts were retained. Label distributions in both gold standards are shown in table 1. Following conventional ML practices, dataset splits are randomly partitioned into a ratio of 3:1:1 for training, validation and testing respectively; while maintaining a stratified set of labels. In total, the union-based dataset contains 3 788 unique segments and the majority-based one comprises 3 571 unique segments[9]. The latter has less segments due the 217 paragraphs that were eliminated because no expert agreement was reached.

In case of multi-label classification, it is not clear which average (macro or micro) best defines a model's performance. As Sebastiani argues, there is no agreement to choose between micro- and macro-averages in literature [18]. Some studies claim that macro-average is fair in case of class imbalance, since all the categories have the same weight, whereas micro-average favours methods that just correctly predict the most frequent categories [22]. However, others (the majority) believe that when the label distribution is not balanced, computation of micro-average is preferable, because a

[8] Here, we only consider high-level categories.

[9] All splits are available for further experiments. See footnote 12.

Labels	Majority-vote gold standard						Union-based gold standard					
	CNN		BERT		BERT-fine-tuned		CNN		BERT		BERT-fine-tuned	
	V	T	V	T	V	T	V	T	V	T	V	T
First Party Collection/Use	83	82	87	88	88	91	83	81	83	84	87	86
Third Party Sharing/Collection	84	82	86	85	87	90	80	79	79	82	83	86
User Access, Edit & Deletion	80	70	82	63	77	73	56	45	54	49	56	65
Data Retention	43	40	42	33	54	56	36	48	36	68	62	71
Data Security	76	75	87	82	87	80	66	72	71	80	73	76
International/Specific Audiences	96	82	94	81	95	83	89	92	87	93	92	92
Do Not Track	91	100	80	100	80	100	80	60	80	60	100	92
Policy Change	80	88	80	88	85	90	69	77	75	78	77	80
User Choice & Control	77	72	75	81	78	81	66	64	64	63	66	65
Introductory/Generic	63	73	75	76	78	79	63	65	74	68	73	67
Practice Not Covered	8	13	18	32	35	35	41	37	44	46	45	48
Privacy Contact Information	86	84	79	80	79	78	79	71	75	71	83	78
Macro Averages	72	71	74	74	77	**79**	67	65	68	70	75	**76**
Micro Averages	79	78	81	82	83	**85**	72	70	73	74	77	**77**

Table 2: F1 for three models on the two gold standards in (%) with tuned epochs on validation; V:Validation; T:Test; Threshold=0.5

micro-average aggregates the contributions of all classes to compute the average metric [20,12]. In order to establish a firm foundation, we report both averages.

Table 2 presents F1 scores across all labels with a threshold equal to 0.5 for the two gold standards. For CNN, we applied Adam with default parameters and with 50% dropout just before the last linear layer (learning rate = 0.001, decay rates: β_1 = 0.9, β_2 = 0.999). *BERT* is optimized with the default configuration and LAMB optimizer [25].

In total, 6 experiments were carried out. The scores obtained (micro-averages ranging from 70-85% and macro-average in range of 65-76% for both gold standards) are considered very accurate, especially in the context of the Fleiss expert agreements, reported in [23], which showed human agreement between 49-91% for the same classes here considered. As expected, for all 6 experiments, micro- outperform macro-averages, because for a few labels, the model is not able to learn the class weights properly due to sample scarcity. For instance, *Data Retention* corresponds to only 2-3% of dataset, and yet this class has 1/12 weight in macro-average calculation; whereas micro-average considers dataset heterogeneity and decreases the impact of scarce categories on the final result. Furthermore, the category *Practice Not Covered* shows low F1 on both gold standards. This category refers to all practices that are not covered by other 11 categories and therefore represents a broad range of topics. Consequently, due to diversity of vocabulary, it is difficult for the model to learn this specific class.

Table 2 shows that even $BERT_{BASE}$ achieves state-of-the-art and further improves the results (without domain-specific embeddings). This is due to the facts that 1) transformers scale much better on longer text sequences because they operate in a concurrent manner; 2) *BERT* is using WordPiece encoding and therefore it has a dictionary which is hard to have an OOV case with it; and 3) it has been trained on massive amounts of data. Moreover, the fine-tuned $BERT_{BASE}$ with 130K corpus privacy policy has significantly enhanced F1 average on both gold standards[10]. Interestingly, fine-tuned *BERT* has improved macro-average more than micro. It is a proof that exploiting a good lan-

[10] Fine-tuning BERT took 33 hours for 3 epochs on a single GPU. Once it is completed, training the classification model takes only a few hours, depending on the number of epochs.

guage model enables the classification model to learn the weights more properly, even with the scarce number of samples.

In order to compare our result to *Polisis*, we present table 3 which provides macro-averages on the union-based gold standard. As mentioned in section 2, *Polisis* used the union-based dataset to report their results. The average lines in the table represent the macro-average of the metric (precision, recall or F1) in predicting the presence of each label and predicting its absence (the 7$^{\text{th}}$ line in the table - F1 - is also included in table 2).

As shown in table 3, we successfully reproduce *Polisis* findings (although with different splits, which remain unavailable) and further improve the result by 5% compared to the state-of-the-art. However, we believe this type of average is not a fair measure for multi-label classification. As shown in table 2, the fine-tuned *BERT* model has nevertheless significantly enhanced macro-averages (from 65% to 76%) which is not visible in table 3, where the enhancement is limited to 5%.

Measure	**CNN**		**BERT**		**BERT-fine-tuned**	
	V	T	V	T	V	T
Precision	81	81	81	84	81	83
Precision-absence	94	94	94	95	95	95
average	**86**	**86**	**86**	**89**	**88**	**89**
Recall	58	57	60	62	70	71
Recall-absence	97	97	97	97	97	97
average	**78**	**77**	**79**	**80**	**84**	**84**
F1	67	65	68	70	75	76
F1-absence	95	95	95	96	96	96
average	**81**	**80**	**82**	**83**	**86**	**86**

Table 3: Macro averages on the union-based gold standard in (%) with tuned epochs on validation; V:Validation; T:Test; Threshold=0.5

5 Discussion

This paper considers notoriously cumbersome privacy policies and investigates automatic methods to assist end-users in comprehending these contractual agreements. The conducted experiments confirm the feasibility of our approach in reaching this objective. Since we are benefiting from supervised ML, the performance of the generated model highly depends on the training dataset quality. As shown in table 1, there is a huge difference between the two gold standards for the *Practice Not Covered* class. In the union-based dataset 642 segments are categorized as *Practice Not Covered*, whereas the majority-based gold standard only records 129 occurrences. Unsurprisingly, for this specific label, all models trained with the union-based dataset outperform the models which were trained by the majority-based one. In addition, 513 variation for the *Practice Not Covered* category between the two gold standards shows high expert disagreement. This was not evident in the original paper [23], because the authors reported

Fleiss' Kappa on the parent category (*Other*) and there is no information on annotator agreement for its subcategories.

Figure 3 shows an example of disagreement on *Practice Not Covered* category in two gold standards. The shown paragraph explains Amazon's policy on treating children's data. In the union-based dataset this segment is annotated with *International and Specific Audiences* and *Practice Not Covered* classes, whereas in the majority-based, it is only labeled with *International and Specific Audiences*.

> **"** [. . .] Amazon.com does not sell products for purchase by children. We sell children's products for purchase by adults. If you are under 18, you may use Amazon.com only with the involvement of a parent or guardian. [. . .] **"**
>
> – *International and Specific Audiences*
> – *Practice Not Covered*

Fig. 3: Disagreement example for the Amazon privacy notice

Regarding label-specific performance, almost all models perform quite well on *Do Not Track* class in spite of the low sample occurrence. This is probably due to a smaller set of terminology that is often used in such paragraphs, including specifically the word *track*. Furthermore, as mentioned earlier, the best human agreement was also achieved on *Do Not Track* class with Fleiss' Kappa equal to 91%, which indicates that our ML models simulate human thinking fairly.

The *BERT* model proves that a good language model achieves high performance even on a domain-specific dataset. It also shows that there is a huge potential to improve the results by fine-tuning the language model with domain vocabularies.

In summary, OPP-115 has proven to be a small, yet reliable dataset for supervised privacy policy classification. However, our experiments confirmed legal text subjectivity for a few classes. One possible solution is decomposing those categories into less controversial subclasses with higher experts agreement. In the above example (Figure 3), breaking the *Specific Audiences* segment into more specific classes will make annotations less subjective, for human experts and machines alike.

To the extent of our knowledge, this is the first effort to establish a standard benchmark on privacy policy classification. In the light of recently enforced data protection laws in the EU, all parties that use and collect personal information must ensure their compliance with `GDPR`. Although OPP-115 consists of policies defined by American companies, most of the top-level categories can still be largely mapped to `GDPR` articles. For instance, the category *First Party Collection/Use* can reflect many practices stated in the Article 13, `'Information to be provided where personal data are collected'` and *User Access, Edit & Deletion* can be linked to Articles 16 & 17 (`'Right to Rectification/Erasure'`)[11]. The OPP-115 dataset also contains annotations at attribute level. By extracting these values from an arbitrary privacy policy, it is possible to perform an in-depth analysis and assist experts to check

[11] Website privacy policies in European union depend also on Directive 2002/58/CE

compliance of privacy policies text based on GDPR. Thus, the approach presented in this paper is a valuable initial step towards compliance checking of privacy policies.

6 Conclusion & Future Work

In this paper we investigate the potential of automatic classification of consent agreements in privacy policy consent forms that are frequently faced by lay users. Our findings are based on the compilation of two gold standards, thus providing a reference privacy policy classification baseline for the relevant research community. To the best of our knowledge, this is the first effort towards a standardized benchmark for privacy policies experiments. The evaluation shows that our best model yields F1 score highs of 77-85% (micro-avg) and 76-79% (macro-avg) for union-based and majority-based gold standards, respectively. Both metrics outperform the reported state-of-the-art. In light of human annotator agreement levels achieved for the same data and classes (ranging from 49%-91%), the results can safely be considered as successful.

The approach and method presented are completely reproducible and all resources and data splits are openly accessible[12]. Since the context surrounding our methods (including the data splits) are available, they can be used as a benchmark for other approaches exploring machine-assisted privacy policy classification for improved human understanding.

In the future, we have identified a number of avenues that can yield further contributions. To further improve the F1 scores achieved, the imbalanced label distribution of OPP-115 (see table 1) could be addressed. A possible solution is to use a weighted objective function with respect to the frequency of the labels. Another approach in consideration is to use sampling techniques to improve the balance. Finally, alternative novel methods can be investigated to take fuller advantage of the three different expert annotations available. In this regard, we will examine the usage of methods (e.g. ensemble) that take the varying labels collectively into consideration.

In conclusion, we intend to continue building upon the baseline achieved and the positive results presented in this paper. As demonstrated by the EU-wide GDPR implementation, data regulation is increasingly recognized as a critical area at a political and governance level, whose impact is felt by all digitally-enabled world citizens. Therefore, although not novel, the application of AI techniques to this area has renewed relevance, and there is great value in exploring automation to support private users entering contractual agreements to have a clearer and more secure understanding of their rights, risks and implications.

Acknowledgment

This work has been partly supported by by the European H2020 project "DAPSI" under the Grant Agreement 871498.

[12] A *supplementary archive* is available online for download: <https://github.com/SmartDataAnalytics/Polisis_Benchmark>. The archive contains *inter alia* the source-code required to reproduce all the experiments, some useful documentation and necessary datasets.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
2. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. pp. 160–167. ICML '08, ACM, New York, NY, USA (2008). https://doi.org/10.1145/1390156.1390177, http://doi.acm.org/10.1145/1390156.1390177
3. Costante, E., Sun, Y., Petković, M., den Hartog, J.: A machine learning solution to assess privacy policy completeness: (short paper). In: Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society. pp. 91–96. WPES '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2381966.2381979, http://doi.acm.org/10.1145/2381966.2381979
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Guntamukkala, N., Dara, R., Grewal, G.W.: A machine-learning based approach for measuring the completeness of online privacy policies. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) pp. 289–294 (2015)
6. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. Proceedings of the 27th USENIX Security Symposium (2018)
7. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
8. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/D14-1181, http://aclweb.org/anthology/D14-1181
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
10. Landesberg, M.K., Levin, T.M., Curtin, C.G., Lev, O.: Privacy online: A report to congress. NASA (19990008264) (1998)
11. Libert, T.: An automated approach to auditing disclosure of third-party data collection in website privacy policies. In: Proceedings of the 2018 World Wide Web Conference. pp. 207–216. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2018). https://doi.org/10.1145/3178876.3186087, https://doi.org/10.1145/3178876.3186087
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
13. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. ISJLP **4**, 543 (2008)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013), http://dl.acm.org/citation.cfm?id=2999792.2999959
15. Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: Proceedings of the 24th International Conference on Machine Learning. pp. 641–648. ICML '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1273496.1273577, http://doi.acm.org/10.1145/1273496.1273577
16. Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society pp. 1–20 (2018)

17. Sathyendra, K.M., Schaub, F., Wilson, S., Sadeh, N.M.: Automatic extraction of opt-out choices from privacy policies. In: AAAI Fall Symposia (2016)
18. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (Mar 2002). https://doi.org/10.1145/505282.505283, http://doi.acm.org/10.1145/505282.505283
19. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1555–1565. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/P14-1146, http://aclweb.org/anthology/P14-1146
20. Van Asch, V.: Macro-and micro-averaged evaluation measures [[basic draft]]. Belgium: CLiPS (2013)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
22. Wiener, E., Pedersen, J.O., Weigend, A.S., et al.: A neural network approach to topic spotting. In: Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval. vol. 317, p. 332. Las Vegas, NV (1995)
23. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., 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). vol. 1, pp. 1330–1340 (2016)
24. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
25. You, Y., Li, J., Hseu, J., Song, X., Demmel, J., Hsieh, C.J.: Reducing bert pre-training time from 3 days to 76 minutes. ArXiv **abs/1904.00962** (2019)