Duc Bui*, Kang G. Shin, Jong-Min Choi, and Junbum Shin

Automated Extraction and Presentation of Data Practices in Privacy Policies

Abstract: Privacy policies are documents required by law and regulations that notify users of the collection, use, and sharing of their personal information on services or applications. While the extraction of personal data objects and their usage thereon is one of the fundamental steps in their automated analysis, it remains challenging due to the complex policy statements written in legal (vague) language. Prior work is limited by small/generated datasets and manually created rules. We formulate the extraction of fine-grained personal data phrases and the corresponding data collection or sharing practices as a sequence-labeling problem that can be solved by an entity-recognition model. We create a large dataset with 4.1k sentences (97k tokens) and 2.6k annotated fine-grained data practices from 30 real-world privacy policies to train and evaluate neural networks. We present a fully automated system, called `PI-Extract`, which accurately extracts privacy practices by a neural model and outperforms, by a large margin, strong rule-based baselines. We conduct a user study on the effects of data practice annotation which highlights and describes the data practices extracted by `PI-Extract` to help users better understand privacy-policy documents. Our experimental evaluation results show that the annotation significantly improves the users' reading comprehension of policy texts, as indicated by a 26.6% increase in the average total reading score.

Keywords: privacy policy; dataset; presentation; annotation; user study; usability

DOI 10.2478/popets-2021-0019
Received 2020-08-31; revised 2020-12-15; accepted 2020-12-16.

***Corresponding Author: Duc Bui:** University of Michigan, E-mail: ducbui@umich.edu
Kang G. Shin: University of Michigan, E-mail: kgshin@umich.edu
Jong-Min Choi: Samsung Research, E-mail: jminl.choi@samsung.com
Junbum Shin: Samsung Research, E-mail: junbum.shin@samsung.com

1 Introduction

Under the FTC framework of *Notice and Choice* [1], privacy policies are a binding contract that services, offered through websites or mobile apps, must adhere to. While this framework is accepted in the US and EU [2], it is up to users to read, and give consent to, the privacy policies. Thus, law and regulations, such as GDPR require services, to provide users with transparent and easy-to-read privacy policies [3].

It is desirable to help users understand the terms used in the privacy notices to raise their awareness of privacy. Despite their growing concerns about data collection and sharing [4, 5], users rarely read them due mainly to their legal sophistication and difficulty to understand [6–8]. Hard-to-understand privacy policies can also lead end-users to blind consent or click-through agreements, risking their privacy since clicking an agreement icon on a website is considered as giving consent to the service provider to lawfully collect and process both general and sensitive personal data [3]. Users are more likely to take necessary steps to protect their privacy if they (especially non-technical users) can understand, and are made aware of privacy at stake [9].

The main thesis of this paper is that *automatic extraction and presentation of data practices help users understand privacy policies better*. The data practices comprise the data objects and privacy actions (collection or sharing) performed thereon. We focus on the users who want to understand the privacy practices, and help them comprehend the privacy notices faster and better. Motivating uninterested readers of privacy documents is orthogonal to the theme of this paper.

Prior work on extracting information from privacy policies has several fundamental limitations. First, existing techniques like PolicyLint [10] use information extraction methods that have high precision but low recall to minimize false positives for their detection of policy contradictions. In contrast, our goal is to achieve both high precision and recall rates. Presenting the data practices to help users improve their reading comprehension requires not only high precision but also high recall because a high false positive or false negative rate (i.e., low precision or recall) will lower the users' con-

fidence and even make them abandon the visualization tool altogether. Furthermore, prior work relied on limited datasets which were either generated from a small number of template sentence patterns [10] or created by non-expert *crowdsourced* workers [11, 12]. A template-generated dataset fails to capture complex and flexible grammatical structures and vocabulary of statements in privacy documents. Crowdworkers are not trained to interpret legal documents so their interpretation may deviate significantly from experts' [13]. Finally, prior information extraction methods are commonly based on a fixed set of manually crafted rules [10, 14, 15] or rely on manual analyses [14, 16, 17], which do not scale to the large number of privacy policies for online services, smartphones and IoT products.

To address the above limitations of prior work, we design and implement a fully automated system, called `PI-Extract`, which accurately extracts data objects and distinct data actions performed thereon (collection/not-collection or sharing/not-sharing). We formulate the information extraction problem as a sequence-labeling problem which can be solved by a named entity recognition (NER) model. We create a large dataset of data practices in real-world privacy policies to train a state-of-the-art neural NER model [18] with contextualized word embeddings [19].

`PI-Extract` presents the extracted data objects and actions as data practice annotation (`DPA`) on privacy policy text to reduce users' burdens in reading and comprehending the policy documents. `DPA` highlights phrases to help users easily identify personal data types in the privacy-policy excerpt and provides a short description of data action to help users determine whether the data types are collected/shared or not. Fig. 1 shows an example of `DPA` created by `PI-Extract`. We have conducted an experiment to evaluate the effect of `DPA` on user comprehension, the impact of wrong predictions, and the effect of annotations on the reading effort. The results show a significant improvement in reading comprehension of `DPA` over the plain text version. Effects of wrong predictions on comprehension and effects of annotations on answering time are also evaluated.

This paper makes the following contributions:

- Construction of a large fine-grained dataset of phrase-level regulated personal information types and the data actions performed on them. The resulting corpus (available on GitHub [20]) comprises 30 real-world privacy policies (4.1k sentences and 97k tokens) with 2.6k annotated data practices and achieves a 98.74% F1 inter-annotator agreement. To the best of our knowledge, this is the largest dataset of fine-grained data practices in real-world privacy policies known to date (Section 5).
- A fully automated system, called `PI-Extract`, which extracts data objects and privacy practices performed thereon. `PI-Extract` leverages a neural NER model, with contextualized word embeddings, trained on our large dataset and achieves an F1 score higher than a rule-based approach based on the method of PolicyLint [10] (Section 6).
- A user study of a presentation method called *data practice annotation* (`DPA`), which presents extracted data types and privacy actions as text highlights and annotations to help users understand privacy policies better. An experiment on 150 users showed that the `DPA` significantly improves the users' comprehension of the privacy texts as indicated by a significant improvement (26.6%) of the average total reading score over the plain text version. The majority of participants found our `DPA` very or extremely helpful in their qualitative feedback. To the best of our knowledge, this is the first application and study of the effects of text highlighting and annotation in reading comprehension of privacy-policy texts (Section 7).

2 What is `PI-Extract` for?

Personal data types and data practice extraction are critical steps in privacy policy analysis. Prior work on privacy policy analysis [10–12] includes these extractions in their pipelines. `PI-Extract`'s extraction improvements will facilitate the development and performance enhancement of privacy policy analysis pipelines.

Presentation of extracted personal data objects and data practices as text annotations in privacy policies can be used in two ways. First, it can be **used after an information retrieval (IR) system** to highlight the data practices in short paragraphs which were previously extracted by the IR system. Highlighting search terms in the snippets of search result pages has been widely used by search engines to help users find the relevant results faster [21, 22]. Prior IR-based approaches, such as Polisis [23], present to the users relevant paragraphs from a privacy policy document, but large chunks of raw text are still daunting for users to read through and comprehend. Our visualization helps users search for information of interest in the text snip-

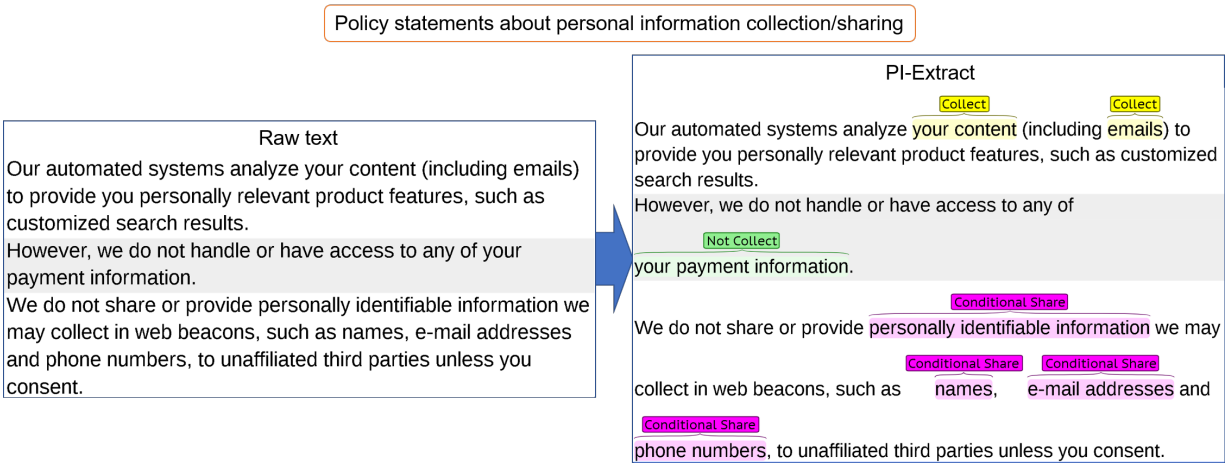


Fig. 1. `PI-Extract` extracts and presents collection and sharing practices of personal information in privacy-policy statements.

pets and read the contextual statements surrounding the phrases of interest.

Second, the presentation can be **used with full privacy policies to facilitate the analysis of non-standardized policies** for researchers, organizations and individuals (such as journalists). For example, `PI-Extract` can be leveraged to assist scientists in recent systematic studies of privacy policies of menstrual apps [24] and mobile money services [25].

3 Related Work

Data Type Extraction. There has been prior work on extracting data types from privacy policies. Costante *et al.* [15] use pattern matching on tokens and named entities to extract personal information types collected by a website. Bhatia *et al.* [16] extract a lexicon of personal information types by identifying noun phrase chunking patterns from 15 human-annotated privacy policies. Bhatia *et al.* [14] and Evans *et al.* [17] use hyponymy patterns to extract personal data types from privacy policies. All of these methods rely on manually-specified rules and lack patterns for extracting data-sharing practices.

PolicyLint [10] extracts the data practices (collection/sharing) on data types to detect contradictions in privacy policies. Its NER model is trained on a small number of samples: only 600 sentences mainly generated from 9 subsumptive patterns, so its dataset and extraction capability are limited in terms of grammar and vocabulary. In contrast, our models are trained on a much larger and more comprehensive dataset — 4.1k sentences (97k tokens) from 30 real-world privacy documents — and thus covers a wider range of grammar and vocabulary. Furthermore, PolicyLint focused on extraction precision (similar to a linter tool), and hence did not evaluate the recall while `PI-Extract` balances between precision and recall to provide users with both correct and complete recognized data types in a document. Therefore, it is not designed to use for helping users understand the text because a low recall rate will provide users with incomplete information and will even reduce the user's confidence in the extraction tool.

GUILeak [12] extracts the data types collected by the services either via user inputs or automatic tools to detect violations in the data collection practices of Android apps. Salvin *et al.* [11] extract from privacy policies the platform information types collected by Android apps and map them to the corresponding Android API functions to detect violations in the implementation of the apps. They only consider data-collection practices, i.e., they do not distinguish data collection from 1st and 3rd parties. PoliCheck [26], built upon PolicyLint [10], can distinguish the receiving entities (1st or 3rd party) when detecting dataflow-to-policy inconsistencies, but suffers from the same limitations of PolicyLint.

Privacy Policy Datasets. Recently, researchers have devised labeled datasets to facilitate the development of machine learning algorithms for automated analysis of privacy policies. OPP-115 [27] is a corpus of annotated paragraphs of 115 website privacy policies. The annotation scheme consists of ten data practice categories, such as 1st-party collection or use, and each data practice has a list of attributes such as data type and purpose. Opt-out Choice dataset [28] includes opt-out choices, such as opt-outs from behavioral advertising. Polisis Twitter QA [23] is a collection of 120 tweets con-

taining questions about privacy policies, alongside the annotated answers obtained from the corresponding privacy policies. APP-350 dataset [29] provides annotated sentences and paragraphs of 350 Android privacy policies, while PI-Extract has finer-grained annotations at the phrase level. Prior datasets are coarser-grained and less diverse than ours, or created by non-expert annotators. They comprise long text spans [27–29], large text segments [23], rigid examples generated from a small set of only 16 patterns [10], or annotations created by non-expert *crowdsourced* workers [11, 12].

User Interfaces for Privacy Policies. Numerous approaches have been proposed to make privacy policies more accessible to users. Polisis [23] retrieves and presents policy paragraphs relevant to a user's question in a chatbot. Since Polisis is based on coarse-grained annotations in OPP-115 dataset [27] at the paragraph level, it can only classify and rank segments of privacy documents. Therefore, PI-Extract can extract data objects at the word and phrase levels while Polisis does not. Moreover, PI-Extract can be integrated with Polisis to enhance the user's understanding of privacy documents further. For example, Polisis can be used to extract the paragraphs relevant to the user's query, and then use PI-Extract to highlight the important phrases about data objects and practices in the paragraphs.

Many researchers worked on various aspects of evaluation and presentation of privacy policies. Disconnect [30] introduces a set of icons to represent privacy risks of a privacy policy. Privacy Nutrition labels [31] present lengthy privacy policies in a nutrition-label-like form. Kay *et al.* [32] show that the visual elements, such as factoids, vignettes, iconic symbols and typography, increase the attention and retention of the users when reading the software agreements. Other research [33, 34] uses a comic-based interface to draw users' attention to privacy notices and terms of service agreements. [35] evaluates three formats for privacy policies and found that the standardized presentations are not effective in helping users understand companies' privacy practices.

4 Background and Problem Formulation

4.1 Neural Named Entity Recognition

Named entity recognition extracts such entities as names of people and places, is commonly formulated as a sequence labeling problem, and then solved by Recurrent Neural Networks (RNN) [36]. RNN encodes the text sequentially and can handle long-term dependencies in text while bi-directional long short-term memory (BLSTM) is one of the most widely-used neural architectures for text classification and sequence labeling [18, 37]. In entity recognition, since the label of each token depends on the probability of its neighbors, a conditional random field (CRF) layer is commonly used after the RNN layer to improve the prediction performance [18].

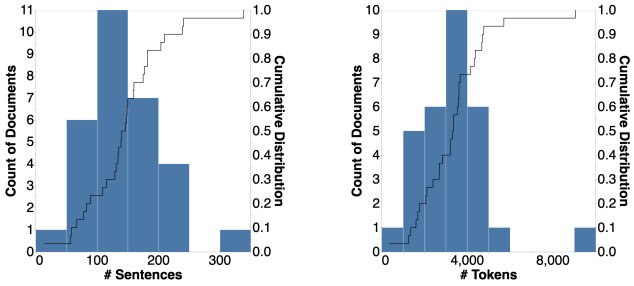
Raw text tokens are converted to real-value vectors before inputting to neural networks by using word embeddings, which comprise the mappings from each word to a single vector. Word embeddings are trained on large datasets of billions of tokens to maximize the coverage of linguistic phenomena. Early word embeddings, such as word2vec [38] and GloVe [39], map words to vectors without context. Recent advances in NLP and computation introduced contextualized word embeddings, such as ELMo [40] and BERT [19], in which the surrounding words are taken into account when mapping a word to a vector, hence improving the prediction performance.

4.2 Problem Formulation

We formulate the extraction of personal data objects and actions thereon as a sequence labeling problem: given a sentence of tokens $s = t_1, \ldots, t_n$, find the label l_i for each token t_i, where $l_i \in \{\textit{Collect}, \textit{Not\_Collect}, \textit{Share}, \textit{Not\_Share}\}$. A personal data object is a text span (a phrase or a word) that expresses a type of user data. Each of such text spans is assigned a data-action label which indicates the action thereon. The labels for text spans are actions on data objects, "collection by 1st party" and "sharing with a 3rd party", and whether the action is performed or not. The 1st party is the company/organization that owns the service, and 3rd parties are companies/organizations other than the 1st party. Determining the labels is based on the data flows: *Collect* and *Share* correspond to the data flows to the 1st and 3rd party, respectively. Table 1 shows their definitions. For example, phrase "your personal information" is marked *Not_Share* in "we may not share your personal information with anyone". We use the classic flat entity structure [41] for each label so that text spans with the same label (i.e., same data action) are contiguous and not overlapping. For example, the whole "delivery and address information" is labeled instead of each overlapping phrase "delivery information" and "address information".

Label	Action performed on the data object
Collect	Collected or used by the first party.
Share	Collected by a third party.
Not_Collect	Not collected and not used by the first party.
Not_Share	Not collected by a third party.

Table 1. Types of data actions to extract from text.



(a) Number of sentences. **(b)** Number of tokens.

Fig. 2. Cumulative distributions of document lengths in terms of number of sentences and tokens.

The labels are used independently without assuming their mutual exclusion or implication. For example, *Share* does not always imply *Collect*, when the service allows a third party to collect and analyze the user's *personal data* instead of doing it by itself, such as in "we do not collect any personal data, but we use Google AdMob that can collect and send it to Google." Furthermore, a pair of negated labels can be used for the same phrase when conditional sharing is performed. "Your personal information" is labeled with both *Share* and *Not_Share* in "we do not share your personal information with third parties without your consent." It is worth noting that handling contradictory policy statements (e.g., a data type is stated to be both collected and not collected) is outside the scope of `PI-Extract`.

5 Dataset Construction

While data objects can be extracted using NER models, creating a dataset is challenging because the determination of start and end of data type spans is vague due to the addition of vague words in the sentences. For example, given a sentence "we collect certain information about your location," we can select either *certain information about your location*, *information about your location*, *your location*, or *location*. A state-of-the-art approach [10] opted to use a set of manually-derived patterns to reduce their efforts. This section describes how we created and controlled the quality of a dataset for training and evaluating the performance of NER for extracting data practices.

5.1 Data Practice Dataset Contruction

5.1.1 Document Selection

We selected and annotated 30 documents from the 115 online privacy policies in the OPP-115 dataset [27] which cover a variety of data practices and styles of online privacy policies. Although OPP-115 cannot be used directly for our purpose of training NER, it contains coarse-grained paragraph classifications which were used as the starting point of our annotation process. We chose the policies of the top websites in the US [42] as large service providers tend to have long and sophisticated policies and have higher coverage of the linguistic phenomena in the corpus [43]. The websites comprises various business domains such as social network, search engine, banking and e-commerce. Total number of sentences and tokens are 4.3k and 99.1k tokens, respectively. Each policy has 144 sentences and 3303 tokens on average. The cumulative distributions of the number of sentences and tokens are shown in Fig. 2.

5.1.2 Annotation Scheme and Process

Two annotators labeled the data objects in each sentence with the 4 labels described in Section 4 and created annotation guidelines for annotators to create consistent labels. The labelers were two of the authors: an advanced PhD student and an industry privacy researcher, and both had more than two years of experience in privacy and security research. First, we created a mini-reference from a subset of 12 documents (40% of the corpus) to develop and evaluate an annotation guideline and process. The main principle is to extract noun phrases from the privacy sentences which express a personal data type that is collected, used or shared by the service provider. The annotation guideline explains corner cases such as how to extract data objects from a complex list. We evolved the guidelines to reflect the new phenomena encountered in the documents while inter-annotator agreement (IAA) was continuously measured to give feedback to annotators. Every time the guidelines were modified, we reflected the changes onto the existing annotations. The guideline

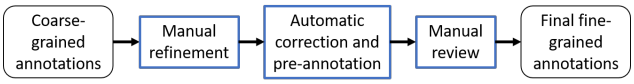


Fig. 3. Semi-automated annotation process.

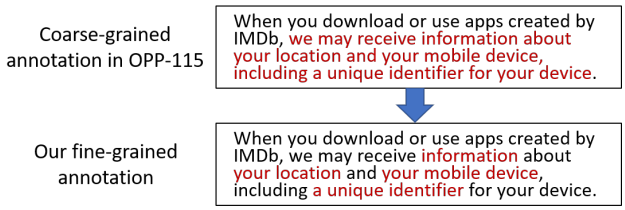


Fig. 4. Example of how long labeled texts in the OPP-115 dataset are refined into shorter phrases. The red color denotes personal information.

document had 4 major updates and its final version (available on GitHub [20]) has 7 pages, 6 high-level principles and 7 rules, each of the rules with multiple examples. After the guidelines and methodology were stabilized and fixed, each annotator followed them to perform the annotation independently on other 18 documents. Finally, they resolved the remaining disagreements by follow-up discussions.

Annotation Revision. To increase the annotation speed and quality (i.e., consistency), we used a semi-automated process that has 4 steps: preprocessing, revision of existing coarse-grained annotations in OPP-15, automated correction/pre-annotation, and final review. These steps were done in sequential order for each document (as shown in Fig. 3). We first removed the sentences which do not contain an actual description of data collection or sharing from the dataset to reduce noisy samples. In particular, we removed sentences which are titles or not a complete sentence. A sentence is considered as a title when it matches the corresponding title-cased statement more than 95% or has less than 4 tokens. The similarity is calculated by using the Levenshtein distance with *fuzzywuzzy* [44] library. Furthermore, since the OPP-115 dataset was in the HTML format, we extracted well-formed plain-text sentences from the HTML, such as merging lists into well-formed sentences and aligning annotations between plain text and HTML code.

The annotators created new fine-grained phrase-level annotations based on the existing coarse-grained labeled text spans in the OPP-115 dataset which was created by law experts. The original OPP-115 dataset has a low overall inter-annotator agreement (IAA) of 29.19% F1 on the 4 labels since it was intended to have classified paragraphs rather than labeled text spans. Therefore, we resolved the conflicting annotations, refined the labels which cover long text, and identified additional data objects that the original annotators missed. While having a low IAA, the existing annotations, created by skilled workers, are useful to speed up the process, such as to determine whether or not a sentence contains any data collection or sharing practice.

Our revision of the OPP-115 corpus was done using WebAnno [45] web-based text annotation tool. An example revision is provided in Fig. 4 where a long-labeled text is refined into three shorter annotated phrases. Other sentences which do not end with a period or do not start with an alphabet character are also removed since they are typically sentence fragments resulting from preprocessing.

Automated Correction and Pre-annotation. We developed a semi-automatic process that includes automated tools for correction and pre-annotation, which are commonly used to increase the annotation speed and improve the quality of corpora [43, 46, 47]. The limitations and bias of the automated methods were also written in the annotation guideline for annotators to be aware of them and avoid too much reliance on the automatic annotations. These tools were developed on 12 policies and fixed thereafter. They were then used to double-annotate the remaining 18 documents.

Automatized Correction. The automatized correction has 2 steps to create consistently labeled text spans: (i) remove relative and prepositional clauses, and (ii) align annotations with noun chunks. Although including relative clauses can narrow the scope of a data type, they frequently contain nested noun phrases, so how to determine the end of these clauses is unclear. For example, "your personal information that you entered in the forums on our website" would be revised to "your personal information". If we include the relative clause, it is hard to determine whether the annotated text span should end at *the forums* or *our website*. Therefore, removing the relative and prepositional clauses reduces the inconsistencies of the labeled spans. The labeled text spans are then aligned to noun chunks in each sentence. The noun phrase alignment removes inconsistencies in the text spans because it is challenging and tedious for annotators to remember to include all the adjective and pronoun prefixes such as "other" and "additional". The alignment also automatically determines whether the conjunctions (*and* and *or*) in a list of data objects would be included in the annotation or not. We used the Spacy library [48] to recognize and chunk non-nested noun phrases.

Automated Pre-annotation. We leverage automatic extraction in PolicyLint [10] to reduce the effort of finding new data objects. Although PolicyLint has a low recall rate, its high precision is useful to reduce the correction effort of the annotators. In particular, we use the domain-adapted named entity recognition (NER) and Data-Entity-Dependency (DED) trees trained in the same dataset in PolicyLint to recognize data objects and label the action for each text span. Our modifications to PolicyLint are detailed in Section 6.

Final Manual Review. After the automatized correction and pre-annotation, the annotators manually reviewed the automatically created annotations. Finally, they hold a discussion to reconcile the disagreements between their labeled policies.

5.1.3 Privacy Policy Corpus

The resulting corpus has 4.1k sentences and 97k tokens. The annotators labeled 2,659 data objects in all documents. The exact-match F1 score is used as the IAA metric. This score has been widely used to measure the prediction performance of the NER task [49]. Two labeled spans match only when they have the same boundaries and the same label. One of the annotators is set as the reference and IAA is then computed as the exact match of the other annotator with the referenced person. The IAA was calculated after the final manual review and achieves 98.74% F1 (98.87% precision and 98.61% recall) overall. The IAA does not reach 100% due to the inherent ambiguity in policy documents and different interpretations of the same sentence. The IAA for each document is presented in Table 13 in Appendix F. We spent an average of 1 hour annotating each policy, or 60 hours in total for 2 annotators.

6 Data Practice Extraction

6.1 Automated Extraction Techniques

6.1.1 `PI-Extract`

`PI-Extract` extracts data objects and the data practices by using neural networks which provide more flexibility than the rule-based methods. While rule-based methods rely on the completeness of the list of collection and sharing verbs, neural models leverage the semantics and syntactic knowledge from word embeddings trained on massive corpora. In particular, as described below, `PI-Extract` uses a BLSTM-CRF model based on BERT-Large-Cased contextual word embeddings [19] to achieve the best performance. Below, we describe the design of `PI-Extract` and experiments with different data practice extraction techniques.

In the BLSTM-CNN-CRF architecture [18], the input text is encoded into a dense vector as the concatenation of word embeddings and character-level representations (encoded by a Convolutional Neural Network (CNN)). The embeddings are then inputted to a layer of BLSTM which encodes the sequence in both backward and forward directions. For a given sentence $(x_1, x_2, \ldots, x_n)$ containing n words, an LSTM computes a representation $\overrightarrow{h_t}$, the left context of the word x_t in the sentence. Another LSTM layer computes a representation $\overleftarrow{h_t}$ for the right context. Thus, each word within the sentence is represented as a combination of the left and right contexts, $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. This representation is then fed to a CRF layer to compute the scores of the labels for each input token with dependency on its neighbors.

`PI-Extract` uses 4 BLSTM-CRF-based NER models to predict the 4 labels in any sentence because each NER model can predict only a single non-overlapping label for each token while different labels can overlap, i.e., a token can have multiple labels assigned to it. Each model is jointly trained on each dataset to recognize both the text boundaries of data objects and the privacy actions (like collection or sharing) performed on them. `PI-Extract` uses the maximum likelihood as the loss function so that the training process maximizes the probability of the correct tag sequences [50].

The BLSTM-CRF network has one 100-dimensional bidirectional LSTM layer. We used L2 regularization for the transitions in the CRF layer with $\alpha = 0.01$. The training phase used a batch size of 20 and an Adam optimizer with a learning rate of 10^{-5} and coefficients ($\beta_1 = 0.9$, $\beta_2 = 0.999$). These parameters are similar to those used in [18]. We experimented with two state-of-the-art pre-trained word embeddings: 300-dimensional GloVe [39] and 1024-dimensional contextualized BERT-Large-Cased [19]. GloVe converts each token to a dense real-number vector regardless of its context while BERT leverages the context in the sentence to generate the output embeddings.

Since it is desirable to balance between high precision and high recall for generic use cases, the model is optimized for the F1 score (i.e., the harmonic mean of precision and recall). The training of the neural models ran for a maximum of 100 epochs and stopped early if F1 did not improve after 10 epochs. `PI-Extract` implemented the neural models using the AllenNLP framework [51].

6.1.2 Rule-based Extraction

To create a strong baseline, we implemented a rule-based extraction (RBE) method based on the open-source code of PolicyLint [10]. PolicyLint uses patterns of dependency trees of sentences to extract policy statements as 3-tuples $P = (Entity, Action, Data)$ where *Entity* performs an *Action* (*collect* or *not-collect*) on the *Data*. A data structure called *Data-Entity-Dependency* (DED) tree is used to analyze the dependency tree of the sentence to extract the policy statements. A DED tree represents the relation between a *Data* and a *Entity* in a sentence's dependency tree.

RBE uses a list of phrases for the corresponding parties to determine the role of an Entity (i.e., a first or third party). The list comprises terms subsumed by the first/third-party phrases (Table 2) in the ontologies of PolicyLint [10] and PoliCheck [26]. RBE matches the lower-cased words if the phrase is a pronoun, or matches the lemmas otherwise. For example, "authorized third-party service providers" contains lemmas "service provider", and is hence classified as a third party.

RBE then determines the label for each *Data* text span based on the role of the *Entity* in each simplified policy statement extracted by PolicyLint, which expresses a data flow to the *Entity*. In particular, the label is *Collect* or *Share* for a first- or third-party *Entity*, respectively. The same action verb can have a different label, depending on the *Entity* role. For example, considering "we may share your personal information with third parties" and "you may be required to share your personal information with us," although they use the same verb *share*, the label of "your personal information" is *Share* in the first sentence but is *Collect* in the second case. Examples of label determination are given in Table 11 (Appendix C).

RBE makes several changes to optimize PolicyLint extraction for the PI-Extract dataset. RBE disables a generation rule of PolicyLint which generates a *Collect* label for every sharing verb since we do not assume any implication between the labels (Section 4.2). Furthermore, RBE adds the clausal complement (*ccomp* dependency) to negative sentiment propagation to improve the extraction of negated verbs. Given data objects extracted by PolicyLint, RBE aligns them to noun chunks following our annotation pipeline (Section 5.1.2). On the other hand, the original entity recognition model of PolicyLint is reused because its data-action extraction algorithm was optimized for the data objects extracted by the model.

Party	Phrases
1st party	I, we, us, our company
2nd party (user)	you, visitor
3rd party	third party, affiliate, advertiser, business partner, partner, service provider, parent corporation, subsidiary, sponsor, government agency, other company, other organization, other party, other service

Table 2. Phrases for determining privacy parties.

Label	Split Name	# Positive Sents	# Data Objects
Collect	Training	575	1311
Collect	Validation	192	409
Share	Training	348	552
Share	Validation	144	209
Not_Collect	Training	37	56
Not_Collect	Validation	14	22
Not_Share	Training	58	72
Not_Share	Validation	15	21

Table 3. Dataset statistics. Positive sentences contain at least one labeled data objects.

6.2 Evaluation

6.2.1 Dataset

We randomly divide the dataset into 23 documents (3035 sentences) for training and 7 documents (1029 sentences) for validation. Denoting a *positive sentence* to be the one with at least one labeled text span, the number of positive sentences and data objects of the dataset for each label are given in Table 3. The *Collect* and *Share* labels have the largest number of training instances with 575 and 348 positive sentences, or 1311 and 552 data objects, respectively. *Not_Collect* and *Not_Share* labels have the fewest number of training examples with only 37 and 58 positive sentences, or 56 and 72 personal data phrases, respectively.

6.2.2 Metrics

We compute the precision, recall and F1 score for the exact matches in which a predicted span is considered as true positive only if it exactly matches the golden standard span [49]. Since our goal is to extract and visualize the data objects as complete as possible, maximizing F1 (geometric mean of precision and recall) is more desirable than only maximizing the precision.

6.2.3 RBE Performance

The performance of RBE is shown in Table 4. Since RBE is designed to maximize the precision of recognition, it has low recall and high precision. With train patterns, while the recall rates are only in 27 – 43%, the precision in all of the labels are in 81–100%. The highest precision is 100% for the *Not_Collect* label, and the lowest is 81.34% for the *Collect* label. The overall F1 is 41.81%.

RBE is limited by the pre-specified vocabulary, grammar and extraction rules. Its list of collection and sharing verbs is not complete. For example, the verb list does not include *ask*, so it missed data practices in sentences like "we ask for your name when you register to use certain features." Furthermore, RBE missed data practices in sentences that have complex grammars outside of its 16 training patterns, such as "we may enable our advertisers to collect your location." RBE could not extract *Not_Share* data objects in negative-sentiment expressions that are not included in its negated-verb extraction rules, such as *your email address* in "we may provide your physical mailing address (but *not your email address*) to a postal service." RBE also failed to recognize negative sentiments in semantically-negated statements like "*under no circumstance* do we collect any personal data revealing racial origin."

The performance of RBE improved slightly when it was trained on the positive sentences (i.e., sentences with at least one data object) from training data. RBE learned 616 patterns from 1438 sentences which comprise 560 original PolicyLint samples (86 patterns learned) and 878 unique positive sentences (530 patterns learned) from the PI-Extract dataset. The overall F1 score increases by 2.46% when it uses patterns learned from sentences in the training set so the recall rate is improved with more known patterns. We conjecture this limited improvement to come from the fact that RBE was not designed to learn directly from complex grammars in the real-world sentences but rather from sentences with simple building-block patterns.

	Without Train Patterns			With Train Patterns		
Label	Prec.	Rec.	F1	Prec.	Rec.	F1
Collect	83.19	24.21	37.50	81.34	26.65	**40.15**
Share	81.69	27.75	41.43	82.43	29.19	**43.11**
Not_Collect	100.0	18.18	30.77	100.0	27.27	**42.86**
Not_Share	100.0	42.86	**60.00**	90.00	42.86	58.06
Overall	83.74	25.72	39.35	82.59	27.99	**41.81**

Table 4. Prediction performance of RBE method. In *With Train Patterns* configuration, RBE was trained on the positive sentences in the training set, in addition to the original PolicyLint samples.

Word Embeddings	Label	Precision	Recall	F1
GloVe	*Collect*	65.78	54.52	59.63
GloVe	*Share*	44.17	43.54	43.86
GloVe	*Not_Collect*	77.78	31.82	45.16
GloVe	*Not_Share*	55.56	47.62	51.28
	Overall	57.87	50.08	53.69
BERT	*Collect*	64.46	69.19	**66.75**
BERT	*Share*	65.82	49.76	**56.68**
BERT	*Not_Collect*	100.0	50.00	**66.67**
BERT	*Not_Share*	72.73	76.19	**74.42**
	Overall	65.71	62.63	**64.14**

Table 5. Prediction performance of neural methods.

6.2.4 PI-Extract Performance

Since the neural models are more flexible than the rule-based methods of PolicyLint, they have higher overall performance (F1) but lower precision. The neural networks leverage the syntactical and vocabulary knowledge in word embeddings which were trained with very large datasets [52]. The contextualized embeddings in BERT have better performance than the traditional embeddings in GloVe. Our evaluation results are summarized in Table 5. When using BERT, the overall F1 score is 64.14%, and F1 is improved 7.1–23.1% across labels, compared with the neural models with GloVe word representations.

Using BERT, the extraction works best on the *Collect* label at 66.75% F1 and worst on the *Not_Collect* label at 56.68% F1. This reflects the recognition accuracy is proportional to the dataset size: *Collect* has the most number of training examples (1311 text spans) while *Not_Collect* has the least (56 text spans). A main reason for the low F1 score is that the vagueness and sophistication of the language used in privacy documents make it difficult to determine the text spans and the actions on them. Since the models with BERT embeddings outperform both GloVe-based configurations and RBE by large margins in all labels, we henceforth use BERT-based models for PI-Extract unless stated otherwise.

Since low recall rates are shown to make a bad impact on the usability of visual presentation of data practices (Section 7), we tried to improve the recall rates of the BERT models by changing the early stopping criterion to stop the training when the *recall rate* did not improve for 10 epochs. However, there is a trade-

off between recall and precision. While the overall recall was improved by 3.03%, the overall precision decreased by 4.67% and F1 reduced by 0.88% (Table 12 in Appendix D). Therefore, to make the model to be generic for a wide range of applications rather than being application-specific, we kept the above models with the higher F1.

6.2.5 Extraction of Context-free Data Objects

We hypothesize that the low F1 scores of the models were due to the limitation of NER models which were designed to extract context-free named entities rather than context-dependent data objects and practices. We test the performance of NER models to extract context-free data objects without the data actions. We derived a set of data object entities by merging all the data action labels into a single *Data* label. In the preprocessing step, sentences without any data collection/sharing verbs (list of such verbs are from [10]) were removed. Overlapping labeled text spans were resolved by keeping the longest text spans. This dataset has 1,737 sentences, 55.3k tokens and 1,736 entities. The corpus was then split into a training set (1,274 sentences, 39.4k tokens and 1271 entities) and test sets (463 sentences, 15.9k tokens and 465 entities). On the test set, the BERT-based NER model achieved an F1 score of 80.0% (79.2% precision and 80.9% recall). This result provides supports that context-free data objects can be extracted with high accuracy by the NER models and the consistency of the annotations on data objects in our corpus.

We developed a rule-based string matching baseline that matches data objects based on the lemmas of all the data-object terms in the training set. This method has an F1 score of 48.65% with 34.37% precision and 83.19% recall. The recall rate does not reach 100% because the validation set still contained unseen terms such as those that were specific to the type of the service (such as *photograph*) and did not occur in the training set. Furthermore, the training set did not include complete combinations of word forms such as it included *personally identifiable information* but not *personally identifying information*. The precision is low because this method does not distinguish the semantics of sentences. For example, a data object can be used in data usage purpose clauses that do not express data collection or sharing practices, such as the service uses encryption "to prevent unauthorized persons from gaining access to *your personal information*."

6.2.6 Performance on Homogeneous Privacy Policies

We evaluate PI-Extract on a homogeneous collection of privacy policies that contains policies of services in the same domain. We hypothesized that PI-Extract would have better performance on such policies since they share a similar vocabulary of data objects. Specifically, we selected 11 policies of news websites from the PI-Extract dataset (listed in Table 13 in Appendix F) to trained the BERT models (described in Section 6.1.1) using the k-fold cross-validation strategy. Each of the 11 policies was held out once to create a dataset such that the validation set comprises the held-out policy and the remaining 10 privacy policies constitute the training set. PI-Extract achieved an average F1 score of 69.56% (79.21% precision and 62.42% recall) which is 5.42% higher than that on the heterogeneous PI-Extract dataset. This result indicates PI-Extract performance can be improved further by training on a dataset in the same domain as the target application.

7 Visual Presentation of Data Practices

7.1 Presentation Method

We propose a presentation method, called *data practice annotation* (DPA), to highlight and describe the data practices extracted by PI-Extract in order to enhance users' understanding of privacy policies. In particular, from the predictions of PI-Extract, the personal data objects are highlighted, and actions performed on the data objects are described as text annotations. The data action labels are displayed on the top of the highlighted phrases so that they do not hinder the reading flow of the users on the policy text. The background colors of the text and labels are different for each label. The presentation is implemented in web browsers using Brat annotation tool [53]. An example is shown in Figure 1.

Although there is a rich body of research on text highlighting [54–58], little has been done on the effects of text highlighting and annotation for user comprehension of privacy policies. Wilson *et al.* [59] found that highlighting relevant privacy policy paragraphs can reduce task completion time and positively affect the perceived difficulty of crowdworkers without impacting their annotation accuracy. However, DPA is different in both granularity and the presentation method. First, DPA annotates policies at a fine-grained phrase level. Second, DPA not only highlights personal data types but also pro-

vides descriptions of privacy practices performed on the data types. The highlighted data objects help users find them faster because the users need not perform a slow linear search through the text since the highlighted text already stands out. The data practice annotation puts explanation of privacy practices into context and helps users read related policy statements easier.

7.2 User Study Design

We design an IRB-approved (Study No. HUM00158893) user study to evaluate the effects of the DPA presentation on users' reading comprehension. The purpose of this experiment is to answer the following questions.

- **RQ1**: If correct data practice annotations are presented, do users understand privacy policy text better, as indicated by a higher total score?
- **RQ2**: If erroneous data practice annotations are presented, do users have worse comprehension?
- **RQ3**: If data practice annotations (which are either correct or incorrect) are presented, do users need less effort to read the policy excerpts, as indicated by shorter answering time?

7.2.1 Subjects

We recruited 150 crowdsourced workers from Amazon MTurk [60] for the survey. All the participants were required to reside in the United States due to restrictions in our IRB. To ensure the participants are experienced, they were required to have a good performance track record which includes a 90%-or-higher task approval rate and at least 1,000 HITs approved. We screened the participants during the training to ensure users have sufficient English skills to read and understand the instructions and privacy statements. The workers spent 9.6 minutes on average (with a standard deviation of 5.6 minutes) to complete the questionnaire. We paid each worker $2.3 so they earned an hourly wage of $14.3 on average, which is higher than the U.S. Federal minimum hourly wage of $7.25 in 2020 [61].

7.2.2 Instruments

We selected 4 excerpts from real-world privacy policies, each of which comprises one or multiple paragraphs. Each excerpt is self-contained and contains coherent content (e.g., anaphoras refer to other words in the same snippet). The privacy policies are of diverse online service types: financial (*wealthfront.com*), gaming (*ea.com*), professional social networking (*linkedin.com*), and virtual private network services (*tigervpn.com*). These types of businesses are known to collect sensitive data about users' finance, children's personal information, social connections, and data transfers. The policies were downloaded as the latest version in August 2020.

Excerpts of privacy policies were presented instead of the whole privacy policies because it is unrealistic for a user to read a thousand-word privacy policy from start to end [62]. We assume users can always narrow down to the sections of their interest by using a table of contents or information retrieval tools like Polisis [23].

We experimented with policy segments of different lengths (short and long) and different difficulty levels (easy and hard) of policy text. There are 4 segments in the study, a combination of two lengths — short and long – and 2 types of highlights — positive and negated. The short paragraphs have 133–184 words (6 sentences) while long paragraphs have 300–349 words (14 sentences). The reading time is expected to be 0.6–1.5 minutes (assuming an average reading speed of 238 words/minute [63]). With 4 excerpts in the questionnaire, the total task completion time for each participant (including answering the demographic survey, training questions and usability questionnaire) was expected to be about 10 minutes.

To evaluate the difficulty of the excerpts, we use Flesch-Kincaid Grade Level (FKG) [64] to measure their readability. FKG computes the average grade a person is expected to completely understand the written text and was used in readability studies of privacy policies [65, 66]. Three incomplete-sentence section titles with 2 words or less (such as "2.1. Services") were excluded to avoid skewing results. The excerpts have an average FKG of 14.32, indicating 14 years of education are expected for full comprehension. This reading difficulty is similar to the average FKG of 14.42 in a recent large-scale privacy policy survey [66]. The easiest policy passage is *linkedin.com* with an FKG of 12.40 and the hardest is the snippet from *weathfront.com* with an FKG of 17.43. Table 6 shows the detailed statistics of the selected policy excerpts.

We used PI-Extract to extract the data practices in the excerpts which were previously unseen by the models. The policy snippets contain 1–19 data practice annotations. All 4 data action labels (Section 5) have at least one occurrence among all snippets. The prediction performance is 71.1% F1 score on average, ranging from 0.6 – 1.0 F1 score. Table 8 provides the number of the data practices and prediction performance for each of the selected excerpts.

	Domain	#Sents (#Words)	FKG	Question (Question Type)	DPA-Err Error Type
E1	*wealthfront.com*	6 (184)	17.43	Q1-1 (Data action)	Omitted annotation
E2	*ea.com*	6 (133)	14.40	Q2-1 (Data action)	Incorrect data action
E3	*linkedin.com*	14 (300)	12.40	Q3-1 (Data action) Q3-2 (Data type)	Incorrect data action Omitted annotation
E4	*tigervpn.com*	14 (349)	13.07	Q4-1 (Data action) Q4-2 (Data type)	Omitted annotation Incorrect data action
	Average	10 (241)	14.32		

Table 6. Domain names, lengths, readability scores, questions and types of annotation errors in DPA-Err version of the selected policy excerpts (E1 – E4).

	Plain (n=52)	DPA-Err (n=49)	DPA (n=49)
Overall	3.69 (1.04)	3.12 (1.07)	4.67 (1.16)
Short Excerpts	1.23 (0.70)	1.49 (0.62)	1.76 (0.48)
Long Excerpts	2.46 (0.90)	1.63 (0.86)	2.92 (0.89)

Table 7. Mean (SD) scores. Max possible total scores in Overall, Short Excerpts and Long Excerpts are 6, 2, 4, respectively. *n* denotes the number of samples.

Excerpt	*Collect*				*Not_Collect*				*Share*				*Not_Share*			
	Prec.	Rec.	F1	Sup.	Prec.	Rec.	F1	Sup.	Prec.	Rec.	F1	Sup.	Prec.	Rec.	F1	Sup.
E1	0.83	1.00	0.91	5	0.00	0.00	0.00	1	0.50	1.00	0.67	1	-	-	-	0
E2	-	-	-	0	1.00	1.00	1.00	1	-	-	-	0	-	-	-	0
E3	1.00	1.00	1.00	13	-	-	-	0	0.75	0.50	0.60	6	-	-	-	0
E4	0.88	1.00	0.93	7	0.00	0.00	0.00	5	1.00	1.00	1.00	2	1.00	1.00	1.00	1

Table 8. Extraction performance of PI-Extract on the 4 policy excerpts. 0% F1 score indicates no prediction made for the label.

Questions. The questions test the comprehension of participants about the content of the excerpts. There are 1 and 2 questions in short and long excerpts, respectively. Multiple-choice questions (rather than yes/no questions) were used to reduce noisy randomly-selected correct answers. There are 2 types of questions: (1) select a correct data action performed on a given data type and (2) select a correct data type given a data action and a condition. In the data action questions, the 4 choices are the 4 data actions as described in Section 5. In the data-type-selection questions, alternatives were created as data types in a similar context to avoid guessing the correct answer without reading. In long excerpts, the first and second questions are based on the facts in the first and second halves of the snippet in that order. While the questions are the same among all excerpt versions, the correct answers are contained in one of the annotations in the DPA version. Table 6 lists the types of questions for each excerpt.

To test a deep understanding of the policy text, the questions include conditions or complex data objects which are referenced across sentences so that the respondents need to read carefully to select the correct answer. For example, one question asks for the data practices on the "personal information from children under 13" which was mentioned and defined in different sentences. The questions and excerpts in the DPA version are listed in Appendix A.

Incorrect Predictions. We created a version (called DPA-Err) of the excerpts which contain incorrect annotations to test their effects on user comprehension. These annotations may occur due to imperfect predictions of neural models used in PI-Extract. We manually injected incorrect annotations by altering the existing annotations which were asked in the questions. There are 2 types of wrong annotations. The first is *omitted annotation* in which the annotation of the data type asked in the question is missing from the excerpt. The second is annotations with an *incorrect data action* label. We consider common wrong predictions of swapping between *Collect* and *Share* labels, and between negated and positive labels (such as *Not_Collect* and *Collect*). Table 6 lists the error types in the DPA-Err version.

7.2.3 Procedures

At a high level, the study follows a between-subject design so that each participant reads one of the versions of the privacy policy excerpts and were asked questions related to their content. The three versions of the policy segments are Plain (raw text), DPA-Err (annotated text with injected errors), and DPA (annotated text with predictions from PI-Extract). Fig. 5 shows the visualization of the process of the user study.

After an initial introduction, the experiment comprises 4 main sections: demographic survey, training,

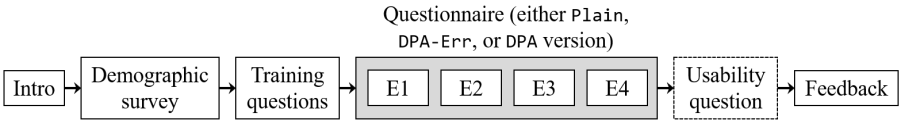


Fig. 5. Visualization of the user study process. Each participant will be shown either Plain, DPA-Err, or DPA version of the policy excerpts (E1–4) in the Questionnaire. Questions in the shaded box are randomly shown to the users. The question in the dashed box is shown only to users of annotated (DPA and DPA-Err) versions.

main questionnaire, and a usability question. The introductory instructions used neutral descriptions without mentioning the annotation presentation in order to prevent participants from forming potential bias. In the main questionnaire, each respondent was presented with either Plain, DPA-Err, or DPA version of the policy excerpts. Questions from the 4 excerpts were also randomly shown to the participants to avoid fatigue effects on a particular excerpt. For each policy snippet, a brief description of the company was provided to the participants to inform them of the context of the privacy statements. We collected the answering time of the participants for each question which was measured from the beginning of the question until the answer was submitted. Due to the limitation of the survey platform which can only measure the submission time per page, participants were shown one question with the corresponding excerpt at a time. The back button was disabled so that participants could not go back to modify their answers.

Since our purpose is to test the reading comprehension, policy excerpts were presented as images to control the results to be only from reading the text, i.e., avoid mixing answers from using a finding tool with answers from reading. Using a finding tool will entail another factor of users' fluency in using the searching tools. To make the text images display consistently among participants, the crowdsourced job description required to perform the questionnaire on a PC or laptop and we programmed the survey to detect the performance on smartphones to terminate the experiment at the first step. The user study was designed and performed via Qualtrics online survey software [67].

Training Questions. Before the main questionnaire, the participants were given two sample questions to help them get used to the main task. Explanations were displayed if they selected wrong answers and they could not proceed until they answered all questions correctly. The instructions also included a notice of the possibility of erroneous data practice annotations due to incorrect predictions.

Usability Question and Feedback. After the main questionnaire, annotated version participants were asked about the usefulness of the annotated text and provided their ratings on a 5-point Likert scale. A final free form feedback form was also provided.

7.3 Experimental Results

We collected a total of 900 responses for the 6 questions from 150 distinct respondents. 52 participants completed Plain, 49 did DPA-Err, and 49 did DPA version. We originally planned to have the same number of workers for each version, but because the participants did the survey simultaneously and some of them left in the middle of the survey, the survey platform did not divide the respondents evenly. All participants completed the survey using a web browser on a desktop operating system and their screens had width and height of at least 1024 and 786 pixels, respectively. In this section, unless noted otherwise, we calculate effect sizes by using Cohen d and the standard deviation is abbreviated as *SD*.

Each correct answer gets 1 score so the maximum possible score of the questionnaire is 6. The score and answering time of each question are shown in Fig. 9 in Appendix B.

7.3.1 Demographics

Across all the respondents, the average age is 45 years (SD=12.1), 49% are males and 50% are females (1% preferred not to answer). 99% of the participants have at least a high school degree (1% preferred not to answer). 41% of the respondents have either high-school education or some college but with no degree while 58% have a bachelor's degree or higher (Fig. 6). 85% of the workers reported being employed.

7.3.2 Research Question 1

The data practice annotations in DPA version improve the reading performance significantly, as indicated by a significant higher total score ($F(1, 99) = 20.06$, $p < .001$, $d = 0.89$). The annotations improve the average total score by 26.6%, from 3.69 ($SD = 1.04$) to 4.67

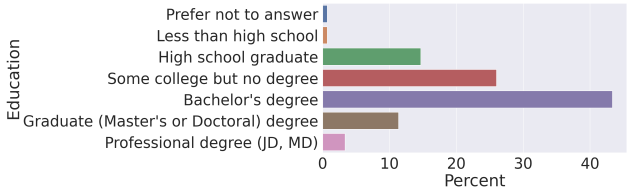
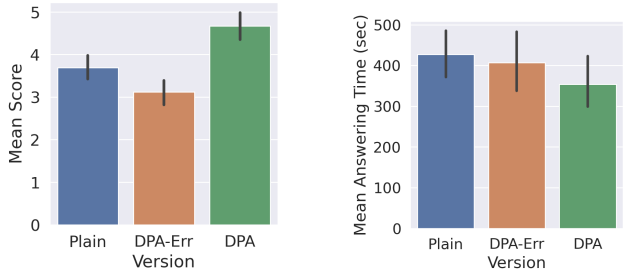


Fig. 6. Education levels of the participants.



(a) Average total scores (max possible total score = 6).

(b) Average total answering time.

Fig. 7. Average total scores and answering time of excerpt versions. Error bars are 95% confidence intervals.

Error Type	Version	Mean (SD)	p-value (d)
Omitted annotation	Plain	1.81 (0.66)	-
	DPA-Err	1.18 (0.67)	$< .001$ (0.94)
Incorrect data action	Plain	1.88 (0.70)	-
	DPA-Err	1.94 (0.63)	0.68 (0.08)

Table 9. Scores on different error types of DPA-Err. The max possible total score of the questions of each type is 3.

Version	Mean (SD)
Plain	427.06 (215.80)
DPA-Err	407.31 (251.54)
DPA	353.97 (226.30)

Table 10. Average total answering time (sec).

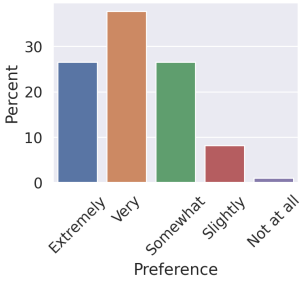


Fig. 8. Helpfulness of annotations (DPA and DPA-Err).

($SD = 1.16$) and the effect size $d = 0.89$ is large [68, 69]. The detailed scores are shown in Fig. 7a and Table 7.

Further analysis shows that the effect of DPA is significant on both short policy excerpts ($F(1, 99) = 18.92$, $p < .001$, $d = 0.87$) and long snippets ($F(1, 99) = 6.63$, $p < 0.05$, $d = 0.51$). The improvement in average total scores of DPA is on short snippets (42.6% increase) which is higher than the long excerpts (18.56% increase). DPA is most effective on the question *Q2-1* which asks about the data action performed on *personal information from children under 13* of *ea.com* with the correct answer to be *Not Collected*. The effect size on this question is large $d = 1.45$ ($F(1, 99) = 53.34$, $p < .001$). We hypothesize that there are fewer annotations in short texts so users spend less time to find the annotations relevant to the question. Table 7 shows the scores on the excerpts.

7.3.3 Research Question 2

The effect on the overall reading performance of wrong annotations in DPA-Err version is significant ($F(1, 99) = 7.35$, $p < 0.01$, $d = 0.54$). The average total score was reduced by 15.43% from 3.69 to 3.12. The effects on short and long excerpts are mixed. While DPA-Err slightly increases the average score by 21.05% ($F(1, 99) = 3.85$, $p = .052$, $d = 0.39$) on short excerpts, it significantly reduces the score on long excerpts by 33.67% ($F(1, 99) = 22.49$, $p < .001$, $d = 0.94$). Table 7 lists the scores.

To identify the causes of the negative impacts of incorrect annotations, we further analyzed the effects of DPA-Err when annotations were either omitted or contained an incorrect data action label. While the reduction of the omission incorrectness on performance is significant ($F(1, 99) = 22.40$, $p < .001$, $d = 0.94$), the decrease caused by incorrect-action-label annotations is non-significant ($F(1, 99) = 0.16$, $p = 0.68$, $d = 0.08$). The omission incorrect type indeed did not add any value to the policy text but action-label-swapped incorrect annotations still helped users find the relevant data types so that they could read the surrounding text to answer correctly. A user reported that s/he "still has to read the sentence, it didn't highlight negatives like *do not... collect*." The detailed scores are listed in Table 9.

7.3.4 Research Question 3

Annotations do not significantly reduce the effort of reading the policy text, as indicated by the shorter average total answering time. The difference of average total answering time among 3 versions (Plain and annotated versions) is not statistically significant ($F(2, 147) = 1.33$, $p = .266$). DPA slightly reduces the average total answering time of Plain version by 17.11% ($F(1, 99) = 2.76$, $p < .10$, $d = 0.33$). The difference of the answering time between DPA-Err and Plain is non-significant ($F(1, 99) = 0.18$, $p = 0.67$, $d = 0.08$). The total answering time is shown in Fig. 7b and Table 10. The answering time for each question is shown in Fig. 9b in Appendix B.

7.3.5 Effect of Education Levels

Since the 4 policy segments have different readability scores, we compute the correlation between the user education levels and the answering scores for each policy excerpt. The results show that users with higher education levels achieved higher scores on the Plain version of excerpt E-1, which requires 17.43 years of education to comprehend and is the hardest in the questionnaire. Specifically, users with a bachelor's degree or higher get a significantly higher average score than the other participants with lower education levels. The average score increases by 36.88% from 0.68 to 0.93 ($F(1, 50) = 6.05$, $p = 0.017$, $d = 0.69$). However, there was no significant difference for other easier excerpts in the Plain version. The average scores were also not significantly different in DPA and DPA-Err versions. We hypothesize that the annotations made the policy excerpts easier to read, thus reducing the difference of scores between education levels.

7.3.6 Qualitative Evaluation

A majority of the participants with the annotated versions (both DPA and DPA-Err versions) found the visual aid helpful. 64.2% of them considered the highlighted text very or extremely helpful while 9.2% considered the annotations provided no or slight help. The DPA version which has relevant annotations was given higher preference: 77.5% of workers considered the highlighted text very or extremely helpful and no participant found the visualization not helpful. Fig. 8 shows the distribution in the annotated versions.

The participants of this study also provided free-form comments which confirm the helpfulness of the visual aids. A participant answering the Plain version said the policies were "still not clear, companies need to be required to do a better job." On the other hand, the DPA "was very effective to find information" and "without the highlights it would take many minutes and much more effort to grasp how complicated this all is."

8 Discussion and Limitations

8.1 Limitations of the Model

PI-Extract is not able to detect implicit data objects and actions which are not stated explicitly in sentences. For example, "if we notice that users in general prefer national political commentary, we might put that content in a special place on the website or in the app" indicates that user preference is collected to promote the political advertisements. However, the model is not able to extract the data and action in such a case. Moreover, personal data types can be mentioned indirectly by referring to other data types in other sentences. For example, in the sentence "when you post comments in response to a story or video on any of our Services, we — and other users — receive *that information*, the phrase "that information" refers to "comments" and requires coreference resolution to extract. These limitations can be alleviated by using more sophisticated natural language understanding techniques that can model and analyze the semantics of implicit statements and analyze privacy policies as a whole, not only on a sentence basis.

The contiguous non-nested entity annotation cannot capture data types in nested or non-contiguous texts such as when multiple data objects are included in a single list. For example, two data objects "software attributes" and "hardware attributes" are included in a complex phrase "software and hardware attributes". Such nested data types can be annotated by using nested-entity annotation scheme [41], but it will require a significantly more complicated annotation scheme. The annotation scheme also does not cover the conditions and purposes of data actions which are left as our future work.

The dataset focuses only on privacy policies on websites and has not explored other platforms such as mobile and IoT devices. However, we observe that it is common for services to have a single privacy policy that covers multiple platforms, especially for popular online services [70]. Therefore, similar data types are used across the policies in different platforms and can be extracted by the PI-Extract models.

Although we hoped NER models can jointly learn to extract personal data objects and the actions performed on them effectively, the overall F1 scores are still low. This is possibly due to insufficient data samples needed for the NER models to learn to distinguish different actions applied to the data types in different contexts. Future advances in natural language processing will improve entity extraction models and require less data, so the performance of PI-Extract will be further improved.

Privacy-policy domain-specific word embeddings trained on large corpora of policies were known to provide performance improvements [23]. However, due to the model complexity, training BERT models on million-policy datasets (such as [29, 71]) would require excessive computation. For example, SciBERT [72]

needed 7 days on an 8-core TPU v3, and BioBERT [73] required 23 days on 8 Nvidia V100 GPUs. We leave the evaluation of domain-specific BERT models as our future work.

8.2 Validity of User Study

Our user study could not fully control the participation of online respondents although we tried to recruit experienced crowdworkers who are more likely to make an adequate effort to complete the survey properly rather than just randomly selecting the answers. However, bias should be reduced because of the between-subject design, random assignment among policy text versions, and the use of multiple-choice questions. It would be better to recruit law experts and interview them to have feedback on the quality of annotation.

The reading environment such as screen resolution was not controlled to be consistent among workers although we tried to enforce the participation via a desktop computer by checking the platform on which the survey was accessed. Furthermore, the study used photos to present to users, preventing them from using the Find tool which is common on browsers. A separate study design to test the effectiveness of the Find tool with DPA is needed because DPA does not require users to know the data objects and data practices in advance while the Find tool is useful only when the user knows the keyword s/he is looking for.

8.3 Limitations and Extensibility of Data Practice Annotation

Similar to the effects of text highlighting which depends on the quality of the highlights and the interaction with the learners, privacy practice annotations improve the user comprehension the most when the predictions are correct and users read the surrounding text to understand the sentence. Text highlighting has been shown to improve user retention if the highlights are relevant to the questions, and vice versa [54–57]. Highlighting could even hurt readers' inference of the text [58].

Wrong predictions from PI-Extract indeed have a negative effect on users, similar to inappropriate annotations which are known to have a harmful effect on reading comprehension [74, 75]. However, even with the presence of the incorrect privacy practice annotations, given annotations with an incorrect data action, users appear to have similar comprehension to the Plain version as shown in the analysis of Research Question 2 (Section 7.3.3). We expect that with more sophisticated models, the prediction accuracy will improve and the wrong predictions will decrease.

More annotated privacy policies would improve the extraction performance of data practices further as the PI-Extract dataset still does not fully cover all the data types and grammatical phenomena. We measured the overall F1 given the validation set (Section 6.2) and the varied sizes of the training sets. The result shows that the F1 score increased with the number of policies (Fig. 10 in Appendix E). The linear regression indicates that, if this linearly increasing trend continued, a training set of 56 policies would be needed to reach the overall F1 of 80%.

PI-Extract annotation scheme and pipeline are generic and can be extended to capture other aspects of privacy policies such as data usage purposes, data retention and opt-out choices. For example, an additional *Usage_Purpose* label can be used to denote the purpose of data collection or sharing. The relation between each data practice and its purposes can be then annotated by link annotations [45].

9 Conclusion

We have sought to automatically extract and present personal data objects and privacy practices performed thereon to help users understand which types of their personal information are collected and shared with third parties in privacy policies. We have constructed a large and fine-grained dataset, based on manual annotations of skilled workers. We have then presented PI-Extract, a fully automated system that uses neural models trained on the corpus to extract data practices from privacy policies and outperforms rule-based techniques. PI-Extract presents the extracted data objects and actions as data practice annotations (DPA) on the policy text. A user study was conducted to evaluate the effect of DPA and incorrect predictions on user comprehension and answering time when reading privacy policy excerpts. DPA made a significant improvement of users' comprehension of the presented policy snippets over the plain text version. The results demonstrate the applicability of PI-Extract in raising privacy awareness and reducing the privacy risks for end users.

Acknowledgement

The work reported in this paper was supported in part by the US National Science Foundation under Grant No. CNS-1646130 and the Army Research Office under Grant No. W911NF-21-1-0057.

References

[1] United States Federal Trade Commission. *Privacy online: a report to Congress*. The Commission, 1998.

[2] OECD, OCDE. The oecd principles of corporate governance. *Contaduría y Administración*, (216), 2004.

[3] European Parliament and Council of the European Union. General data protection regulation. page 88, 2016.

[4] Aleecia McDonald and Lorrie Faith Cranor. Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising. SSRN Scholarly Paper ID 1989092, Social Science Research Network, Rochester, NY, August 2010.

[5] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. pages 77–96, 2016.

[6] F. H. Cate. The Limits of Notice and Choice. *IEEE Security Privacy*, 8(2):59–62, March 2010.

[7] F. Schaub, R. Balebako, and L. F. Cranor. Designing Effective Privacy Notices and Controls. *IEEE Internet Computing*, 21(3):70–77, May 2017.

[8] Florian Schaub. Nobody reads privacy policies – here's how to fix that, October 2017.

[9] D. Malandrino, V. Scarano, and R. Spinelli. How Increased Awareness Can Impact Attitudes and Behaviors toward Online Privacy Protection. In *2013 International Conference on Social Computing*, pages 57–62, September 2013.

[10] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 585–602, Santa Clara, CA, August 2019. USENIX Association.

[11] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D. Breaux, and Jianwei Niu. Toward a framework for detecting privacy policy violations in android application code. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, pages 25–36. ACM, 2016.

[12] X. Wang, X. Qin, M. Bokaei Hosseini, R. Slavin, T. D. Breaux, and J. Niu. GUILeak: Tracing Privacy Policy Claims on User Input Data for Android Applications. In *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, pages 37–47, May 2018.

[13] Reidenberg *et al*. Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding. *Berkeley Technology Law Journal*, 2015.

[14] J. Bhatia, M. C. Evans, S. Wadkar, and T. D. Breaux. Automated Extraction of Regulated Information Types Using Hyponymy Relations. In *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, pages 19–25, September 2016.

[15] Elisa Costante, Jerry den Hartog, and Milan Petković. What Websites Know About You. In Roberto Di Pietro, Javier Herranz, Ernesto Damiani, and Radu State, editors, *Data Privacy Management and Autonomous Spontaneous Security*, pages 146–159, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[16] J. Bhatia and T. D. Breaux. Towards an information type lexicon for privacy policies. In *2015 IEEE Eighth International Workshop on Requirements Engineering and Law (RELAW)*, pages 19–24, August 2015.

[17] M. C. Evans, J. Bhatia, S. Wadkar, and T. D. Breaux. An evaluation of constituency-based hyponymy extraction from privacy policies. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*, pages 312–321, 2017.

[18] Xuezhe Ma and Eduard Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. pages 1064–1074. Association for Computational Linguistics, 2016.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] Duc Bui. PI-Extract Dataset https://github.com/um-rtcl/piextract_dataset.

[21] Hui Zhang. Beyond Query-Oriented Highlighting: Investigating the Effect of Snippet Text Highlighting in Search User Behavior, 2018. ISSN: 1687-5265 Publisher: Hindawi Volume: 2018.

[22] Marti A. Hearst. Search User Interfaces, September 2009. ISBN: 9780521113793 9781139644082 Library Catalog: www.cambridge.org Publisher: Cambridge University Press.

[23] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, Baltimore, MD, 2018. USENIX Association.

[24] Laura Shipp and Jorge Blasco. How private is your period?: A systematic analysis of menstrual app privacy policies. 4:491–510.

[25] Jasmine Bowers, Bradley Reaves, Imani Sherman, Patrick Traynor, and Kevin Butler. Regulators, mount up! analysis of privacy policies for mobile money services. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security*, SOUPS '17, pages 97–114. USENIX Association.

[26] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with policheck. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 985–1002. USENIX Association, August 2020.

[27] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. The Creation and Analysis of a Website Privacy Policy Corpus. pages 1330–1340. Association for Computational Linguistics, 2016.

[28] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. Identifying the provision of choices in privacy policy text. pages 2774–

2779. Association for Computational Linguistics, 2017.

[29] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. 2019, 2019.

[30] Disconnect. Disconnect Privacy Icons, July 2017.

[31] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 4:1–4:12. ACM, 2009.

[32] Matthew Kay and Michael Terry. Textured Agreements: Re-envisioning Electronic Consent. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 13:1–13:13, New York, NY, USA, 2010. ACM.

[33] Bart Knijnenburg and David Cherry. Comics as a Medium for Privacy Notices. In *Proc. USENIX SOUP*, 2016.

[34] Madiha Tabassum, Abdulmajeed Alqhatani, Marran Aldossari, and Heather Richter Lipford. Increasing User Attention with a Comic-based Policy. In *Proc. CHI*. ACM, 2018.

[35] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. A Comparative Study of Online Privacy Policies and Formats. In Ian Goldberg and Mikhail J. Atallah, editors, *Privacy Enhancing Technologies*, Lecture Notes in Computer Science, pages 37–55. Springer Berlin Heidelberg, 2009.

[36] Vikas Yadav and Steven Bethard. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

[37] Christopher Manning. Representations for Language: From Word Embeddings to Sentence Meanings | Simons Institute for the Theory of Computing, March 2017.

[38] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119. Curran Associates Inc.

[39] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[40] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[41] Jenny Rose Finkel and Christopher D. Manning. Nested Named Entity Recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 141–150, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[42] Alexa Internet, Inc. Alexa - top sites in united states - alexa.

[43] Karën Fort. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons, June 2016. Google-Books-ID: n7plDAAAQBAJ.

[44] seatgeek/fuzzywuzzy. original-date: 2011-07-08T19:32:34Z.

[45] Richard Eckart de Castilho, Eva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[46] Martin Zurowietz, Daniel Langenkämper, Brett Hosking, Henry A. Ruhl, and Tim W. Nattkemper. MAIA—A machine learning assisted image annotation method for environmental monitoring and exploration. *PLOS ONE*, 13(11):1–18, 2018.

[47] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, June 2009.

[48] Explosion AI. Models & Languages · spaCy Usage Documentation, 2020. Library Catalog: spacy.io.

[49] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[50] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[51] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. Association for Computational Linguistics.

[52] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs]*, February 2020. arXiv: 2002.12327.

[53] Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. Bionlp shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[54] Lucy Cui. MythBusters: Highlighting helps me study.

[55] Robert L. Fowler and Anne S. Barker. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, 59(3):358–364, 1974.

[56] Sherrie L. Nist and Mark C. Hogrebe. The role of underlining and annotating in remembering textual information. 27(1):12–25. Publisher: Routledge.

[57] Jay Blanchard and Vincent Mikkelson. Underlining performance outcomes in expository text. 80(4):197–201.

[58] Sarah E. Peterson. The cognitive functions of underlining as a study technique. 31(2):49–56. Publisher: Routledge.

[59] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. Crowdsourcing annotations for websites' privacy policies: Can it

really work? In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 133–143. ACM Press, 2016.
[60] Amazon Mechanical Turk, Inc. https://www.mturk.com/.
[61] Minimum Wage | U.S. Department of Labor. https://www.dol.gov/agencies/whd/minimum-wage, 2020.
[62] A.M. McDonald and L.F. Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:540–565, 2008.
[63] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. 109:104047.
[64] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
[65] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 18–25. Association for Computing Machinery.
[66] Mukund Srinath, Shomir Wilson, and C. Lee Giles. Privacy at scale: Introducing the PrivaSeer corpus of web privacy policies.
[67] Qualtrics. Online survey software https://www.qualtrics.com/, 2020.
[68] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, 2nd ed edition.
[69] Shlomo Sawilowsky. New effect size rules of thumb. 8(2).
[70] Stephanie Winkler and Sherali Zeadally. Privacy policy analysis of popular web platforms. 35(2):75–85. Conference Name: IEEE Technology and Society Magazine.
[71] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. *arXiv:2008.09159 [cs]*, September 2020. arXiv: 2008.09159.
[72] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.
[73] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. Publisher: Oxford Academic.
[74] Vicki L. Silvers and David S. Kreiner. The effects of pre-existing inappropriate highlighting on reading comprehension. 36(3):217–223.
[75] Vicki Silvers Gier, David S. Kreiner, and Amelia Natz-Gonzalez. Harmful effects of preexisting inappropriate highlighting on reading comprehension and metacognitive accuracy. 136(3):287–302. Publisher: Routledge.

Appendix A User Survey Instruments

The following is the questions used in the `DPA` version of the survey described in Section 7. The `Plain` version is the same except does not have the highlighted text while `DPA-Err` version has annotations which are ommitted or contain an incorrect action label.

[Introduction]
We would like to understand your opinion about the presentation of privacy policies of websites.
By continuing you agree with the collection of your answers in the survey. Your responses for this survey are used for academic research purposes only.
The survey will take 5-10 minutes to complete.

[Demographic Questions]
What is the highest level of school you have completed or the highest degree you have received?
- ○ Less than high school degree
- ○ High school graduate (high school diploma or equivalent including GED)
- ○ Some college but no degree
- ○ Bachelor's degree in college (4-year)
- ○ Graduate (Master's or Doctoral) degree
- ○ Professional degree (JD, MD)
- ○ Prefer not to answer

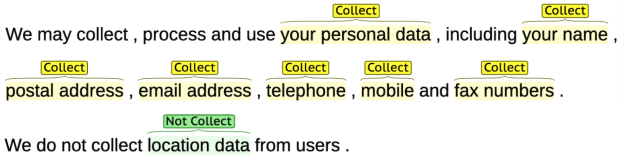
What is your gender?
- ○ Male
- ○ Female
- ○ Prefer not to answer

Are you employed?
- ○ Yes
- ○ No
- ○ Prefer not to answer

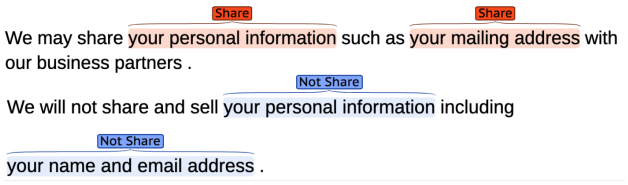
What is your year of birth? [**A text box is presented**]

[Training Questions]

To help you understand privacy policies faster, the following sentences highlight the data that the company collects or does not collect from users.



To help you understand privacy policies faster, the following text highlights the user's data that the company shares or does not share with other businesses.



Read the sentence and answer the questions about a company's privacy policy. You can leverage the highlights to answer faster. Note that some highlights sometimes may be missing or contain an incorrect label.

May the company collect your personal information?
- Yes
- No

May the company share your e-mail address?
- Yes
- No

[**Main Questionnaire**]
[**Excerpt E1**]
Read the following paragraph from privacy policy from a financial service and answer the question below.

Share
Many of Wealthfront 's Users and Clients choose to aggregate information from accounts at other financial institutions onto their dashboard on our Site or in our App ; in enabling this functionality , Wealthfront acts as an agent to retrieve the User or Client account information maintained by such third - party financial institutions with which the User or Client has a legally - binding customer relationship (" Account Information ") .

Collect Collect Collect
This Account Information may include account balances , transactions and holdings from the linked financial institutions .

Collect
By choosing to use our Services to aggregate and analyze your Account Information , you expressly authorize and direct Wealthfront , on your behalf , to electronically retrieve Collect all Account Information associated with the username and password that you use to link the account .

Collect
Wealthfront does not store login credentials used to link Account Information .
Rather , Wealthfront works with one or more third - party service providers to access and retrieve your Account Information .

Share
Any Account Information that Wealthfront receives is read - only and can not be altered by Wealthfront or the third - party service provider we use to access and retrieve your Account Information .

As stated in the paragraph, which of the following practices is true about your transactions from linked financial instutions?
- Collected by the service
- Not collected by the service
- Shared by the service
- Not shared by the service

[**Excerpt E2**]
Read the following paragraph from the privacy policy of a gaming service and answer the question below.

Children
Protecting children 's privacy online is extremely important to EA .
Many EA online and mobile games and Services are intended for adults and do not knowingly collect
Not Collect
any personal information from children .
Meanwhile , other Services provide a different experience for players based on their age .
When players identify themselves as being children we will : (1) not provide a path for them to share personal information , (2) collect certain information for limited purposes only , (3) block or restrict the child from accessing relevant Services , such as chat functionality ; and/or (4) obtain consent from parents for the use of their children 's personal information , all according to applicable law .
When we say children , we mean under the age of 13 in the United States , or the minimum age in the child 's territory .

As stated in the paragraph, which of the following practices is true about personal information from children under 13 in the United States?
- Collected by the service
- Not collected by the service
- Shared by the service
- Not shared by the service

[**Excerpt E3**]
Read the following paragraph from the privacy policy of a professional social network and answer the questions below.

2.1 Services
Our Services help you connect with others , find and be found for work and business opportunities , stay informed , get training and be more productive .

Collect
We use your data to authorize access to our Services and honor your settings .
Stay Connected
Our Services allow you to stay in touch and up to date with colleagues , partners , clients , and other professional contacts .
To do so , you can " connect " with the professionals who you choose , and who also wish to " connect " with you .
Subject to your and their settings , when you connect with other Members , you will be able to search each others ' connections in order to exchange professional opportunities .

Collect
Share
Collect
Collect
Collect
We use data about you (such as your profile , profiles you have viewed or data provided through address book uploads or partner integrations) to help others find your profile , suggest connections for you and others (e.g. Members who share your contacts or job experiences) and enable you to invite others to become a Member and connect with you .

Collect
Collect
You can also opt - in to allow us to use your precise location or proximity to others for certain tasks (e.g. to suggest other nearby Members for you to connect with , calculate the commute to a new job , or notify your connections that you are at a professional event) .
It is your choice whether to invite someone to our Services , send a connection request , or allow another Member to become your connection .

Collect
When you invite someone to connect with you , your invitation will include your network and
Collect
Collect
Share
Collect
Share
Collect
Share
Collect
basic profile information (e.g. , name , profile photo , job title , region) .
We will send invitation reminders to the person you invited .
You can choose whether or not to share your own list of connections with your connections .
Visitors have choices about how we use their data .

As stated in the paragraph, which of the following practices is true about your precise location?
- Collected by the service
- Not collected by the service
- Shared by the service
- Not shared by the service

As stated in the paragraph, which of the following is shared with another person who you invite to connect?
- Your job title
- Your address
- Your professional skills
- Your preferred social networks

[**Excerpt E4**]
Read the following paragraph from the privacy policy of a virtual private network website and answer the questions below.

To provide you with our service we need to authenticate your VPN credentials on our backend .
We need to do this in order to verify that your account is valid and in good standing (active) .
Share Collect
Our backend will also collect data consumption information that is necessary to detect abuse and irregularities connected to our network integrity .
Collect
We store MB values per session (e.g. 895 MB consumed on 01.04.xx) and save that data for 6 months .
We delete that data automatically after 6 months but it helps us understand the bandwidth growth and network integrity over individual time periods and allows our engineers to increase capacity upgrades before reaching a bottleneck .
Collect
At no time , we store , read , analyze or in any other way process the traffic exchanged between you , our servers and the public internet .
In other words , we do not save , read or have technical access to any DNS queries , websites you visit , data you transferred or communications .
tigerVPN sells subscription to pay for its service and has never and will never sell , share , or give
Not Share
away any data .
Share
At tigerVPN customers connect to a VPN server and share the IP address between thousands of other customer connected .
This means that outgoing traffic has the same IP address for every customer at the same time .
We are not able to identify the customer individually because we do not provide exclusive (a dedicated IP addresses per customer) for a VPN connection .
Collect
While we do store a record when you connect to a server (for the sole purpose to provide troubleshooting and accounting , abuse prevention and network integrity) it does not allow us to single out an individual customer because your information overlaps with thousands of other customers at the same time .
Collect Collect
E.g. if Bruno connects with his iPhone to our New York server , a record of that session (start time ,
Collect Collect
end time , data transferred (in MB) is stored in our backend but it does not allow us to single out Bruno as there are thousands of active connections overlapping at the very same time with the very same location .

As stated in the paragraph, which of the following practices is true about your data?
- ○ Collected by the service
- ○ Not collected by the service
- ○ Shared by the service
- ○ Not shared by the service

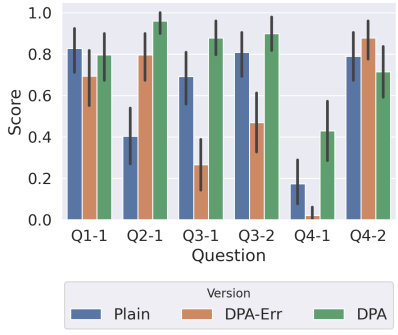
As stated in the paragraph, which of the following may be stored by the service when you connect to a server?
- ○ The IP address you used
- ○ A record of your session
- ○ The messages you sent
- ○ A unique ID of your device

[**Usability Question**]
How do the highlighted words help you identify the personal information collected or shared by the company?
- ○ Not at all helpful
- ○ Slightly helpful
- ○ Somewhat helpful
- ○ Very helpful
- ○ Extremely helpful

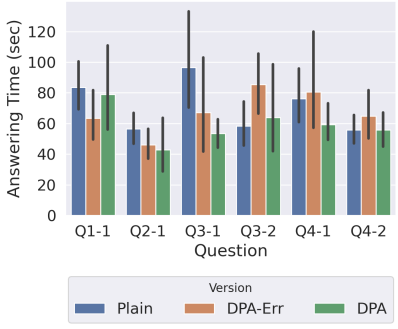
[**Feedback**]
What is your feedback about this survey (if you have)? [**A text box is presented**]

Appendix B Scores and Answering Time

Scores and answering time in user study are shown in Fig. 9.



(a) Score of each question.



(b) Answering time of each question.

Fig. 9. Score and answering time of each question in the user study. Error bars are 95% confidence intervals.

Appendix C Data Action Examples in RBE

Examples of data actions in RBE are given in Table 11.

Entity Role	Data Action	Example
First party	Collect	*We may collect your personal information from Analytics tools.*
Third party	Share	*Our business partners may collect your demographic information.*

Table 11. Examples of data actions, based on simplified policy statements of PolicyLint, used in RBE.

Appendix D Recall-optimized BERT models

The performance of recall-optimized BERT models is shown in Table 12.

Word Embeddings	Label	Precision	Recall	F1
BERT	*Collect*	62.31	71.15	66.44
BERT	*Share*	55.12	54.07	54.59
BERT	*Not_Collect*	77.78	63.64	70.00
BERT	*Not_Share*	76.19	76.19	76.19
BERT	Overall	61.04	65.66	63.27

Table 12. Recall-optimized BERT models.

Appendix E Dataset Coverage

The performance of `PI-Extract` for varied dataset sizes is shown in Fig. 10.

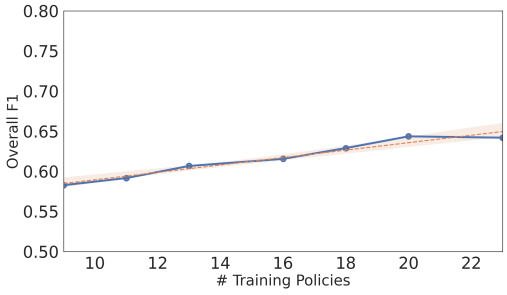


Fig. 10. Overall F1 when increasing the training set size. The linear regression line is dashed and the shade region shows its 95% confidence interval.

Appendix F Corpus IAA and Statistics

The IAA between annotators, number of sentences and tokens of each document in the corpus are shown in Table 13.

Website	Precision	Recall	F1	Support	# Sentences	# Tokens
bankofamerica.com	95.73	91.06	93.33	123	187	4618
yahoo.com	97.83	93.75	95.74	48	76	1573
nytimes.com*	97.96	96.00	96.97	150	200	4317
barnesandnoble.com	97.35	97.78	97.56	225	310	8944
google.com	97.48	98.31	97.89	118	123	3151
instagram.com	97.92	97.92	97.92	96	148	3511
reddit.com	96.83	99.19	97.99	123	163	3536
thefreedictionary.com	100.00	97.30	98.63	37	58	1230
playstation.com	98.68	98.68	98.68	76	135	3484
ted.com	98.41	100.00	99.20	62	54	1336
pbs.org*	100.00	98.48	99.24	66	119	2659
aol.com*	100.00	98.68	99.34	76	135	3291
washingtonpost.com*	100.00	98.73	99.36	79	156	3227
sciencemag.org*	98.77	100.00	99.38	80	128	3195
geocaching.com	100.00	98.78	99.39	82	140	2630
walmart.com	98.84	100.00	99.42	85	228	4589
theatlantic.com*	99.03	100.00	99.51	102	153	4049
gamestop.com	99.12	100.00	99.56	112	169	4295
foxsports.com*	100.00	99.13	99.56	115	126	3590
uh.edu	100.00	100.00	100.00	10	14	343
imdb.com	100.00	100.00	100.00	33	109	2355
thehill.com*	100.00	100.00	100.00	41	53	1669
steampowered.com	100.00	100.00	100.00	56	70	1760
ticketmaster.com	100.00	100.00	100.00	59	147	2054
minecraft.gamepedia.com	100.00	100.00	100.00	73	101	2806
msn.com*	100.00	100.00	100.00	78	86	2090
mlb.mlb.com*	100.00	100.00	100.00	103	122	3606
fool.com	100.00	100.00	100.00	108	183	4734
amazon.com	100.00	100.00	100.00	111	143	3307
esquire.com*	100.00	100.00	100.00	132	228	5700
Total	-	-	-	2,659	4,064	97,649

Table 13. IAA and statistics of privacy policies in the corpus. *-marked websites were used in the evaluation of `PI-Extract` for policies in the same domain (Section 6.2.6).