Contextual Data Type Entity Recognition in Privacy Policies

Natural Language Processing

CSCI-UA 469

by

Vishram Doodnauth
Samuel Yang
Brian Li
Coco Ke

Adam Meyers

Abstract.

Despite the original intent of privacy policies to shed light on data collection procedures, today's privacy policies create significant and time-consuming challenges to end users due to their length, lack of standardization, and dense legal jargon. Their density makes it difficult to identify key data-handling practices within such documents. Consequently, an automated system capable of doing so has numerous applications in aiding user understanding and evaluating compliance with data privacy laws. This paper proposes a context-based NER system to extract four major categories of data practices: information collected by the company, information not collected, information shared with third parties, and information not shared. Through the application and fine-tuning of a BERT token classifier and linguistic post-processing rules, such as negation handling and dependency-based role identification, this hybrid system differentiates between first- and third-party data flows with a token-level F1-score of 0.70, outperforming our manual rule-based baseline and previous neural approaches to this task.

Introduction.

Privacy policies are lengthy and dense legal documents presented alongside computer software that describe how companies and services collect, use, and share the personal data of their users. Despite their importance and relevance, they are notoriously difficult for users to understand. In fact, the average privacy policy contains over 6,000 words and takes nearly half an hour to read (Šlekytė). Even then, the lack of standardization and heavy reliance of the text on legal jargon make it a challenge for users to digest the information they actually care about. Prior work shows that users struggle to interpret privacy policies, often finding themselves misunderstanding or overlooking critical data practices (Brunotte et al.). Consequently, most people tend to ignore privacy policies altogether, which becomes a problem for transparency and compliance.

This work explores the task of automatically recognizing and extracting key data-handling practices from privacy policies. A complete system for understanding privacy practices tackles several problems:

i. It must identify the types of data being discussed, which may appear throughout highly varied and long-winded language.
ii. It must determine whether a particular data type is collected or explicitly *not* collected by the company.
iii. It must detect whether that data is shared or explicitly *not* shared with third parties.

Such problems require interpreting contextual cues, legal disclaimers, and often subtle or vague negative constructions. Prior work on answering privacy practice questions (Ravichander et al.), detecting vague language (Lebanoff et al.), and recognizing third-party entities (Hosseini et al.) has shown that NLP in the privacy policy domain is inherently challenging, yielding modest F1-scores far below human performance.

Objective of the Study.

In this paper, we focus on the task of identifying four key categories of data practices with respect to user information, following the taxonomy described in Bui et al.: data collected by the first party, data *not* collected by the first party, data shared with third parties, and data *not* shared with third parties. Our approach fine-tunes a BERT-based named entity recognizer on a modified version of the Pi-Extract corpus (Bui et al.), a privacy policy corpus consisting of manually annotated mentions of user data and their associated data practices, augmented with a

set of post-processing linguistic rules to recognize text spans denoting personal information. We then categorize such spans under one of the four data practice categories mentioned above. A key challenge revolves around exploiting contextual cues—such as negation, dependency structure, and part-of-speech patterns—to distinguish between first- and third-party **data flows** (i.e., user info recorded solely by the first party vs. user info sent to a third party).

We validate our model through experiments on a manually annotated corpus, comparing against a manual rule-only baseline and past work. A successful privacy practice extraction system can benefit users seeking transparency on data practices, assist regulators in auditing compliance, and enable further downstream analysis of the privacy behaviors of software services.

Related Work.

Many prior studies have attempted to evaluate data practices and transparency through the automated analysis of privacy policies. However, most of these studies classify information found in privacy policies at a broader, more coarse level.

For instance, Ramanath et al. align sections of over a thousand privacy policies through the use of an unsupervised Hidden Markov Model to group large bodies of text in privacy policies under broad categories such as "deletion of personal information." This is done with an F1-score of 0.53. While section-level alignment in privacy policies is useful for analysis across a larger corpus, it remains limiting for the end user, who must still read and interpret dense chunks of text regarding the services they use to find information on the specific data practices and data items that concern them. In contrast, our work targets finer-grained annotations of user data and their associated practices, making it easier for end users and regulators to comprehend individual privacy policies.

Hosseini et al. develop multiple NER models to automatically identify and classify third-party entities found in mobile app privacy policies. They do so by distinguishing between a generic third party (e.g., analytics providers, ad networks) and specific third-party organizations (e.g., Meta, Google Analytics), which is done with an average F1-score of 0.66. Although their model identifies third-party recipients, they do not distinguish which specific data items are sent over to those third parties, nor do they highlight whether certain data types are explicitly mentioned not to be shared. Therefore, this paper directly complements their work by characterizing exactly *what* information about the user is flowing along data edges to the third parties that Hosseini et al. identify. Moreover, Hosseini et al. first developed their system using a newswire-trained CRF model and deduced that effective labeling and extraction necessitate the use of rich linguistic features and context-sensitive modeling in the privacy policy domain. Our approach follows this insight by incorporating linguistically-informed post-processing rules for improved results in identifying data flows tied to user information.

The closest-related approach, authored by Bui et al., automates the classification of personal data flows in privacy policies for use in a user study about privacy policy readability and comprehension. It does so with an F1-score of 0.64. The study demonstrates that visually highlighting data practices (i.e., this data is collected, is not collected, is shared, or is not shared) significantly improves the comprehension of users when reading privacy policies. Bui et al. also conclude that effective automated privacy policy analysis benefits from contextual word embeddings, especially given the limited availability of fine-grained annotations in the domain's current state. Building on this and Hosseini et al.'s findings, our work fine-tunes a BERT-based

model on a privacy policy corpus to leverage contextual word embeddings while also developing linguistically-informed rules to yield higher F1-scores than existing automated systems.

Methodology.

Corpus Selection.

We base our experiments on a modified version of the Pi-Extract corpus introduced by Bui et al., which consists of a collection of expert-annotated privacy policies that make up over 4.1 thousand sentences and 97 thousand tokens. There are about 2.7 thousand labeled phrase-level mentions of personal datatypes (e.g., "email address" or "location information") and associated data flows indicating whether the datatype is collected/not collected/shared/not shared. In the original setup, the same underlying dataset is used to train multiple disjoint sequence-labeling models, each focused on a separate data flow type, effectively creating multiple parallel copies of the data—one per action.

For the purposes of our work, we merge these copies into a single token-level BIO corpus that covers all four practice categories simultaneously. That is, our datatypes are tagged as follows:

- COLLECT: data that the policy explicitly states is collected by the first party.
- NOT_COLLECT: data that the policy explicitly states is not collected by the first party.
- SHARE: data that the policy explicitly states is shared/disclosed to third parties.
- NOT_SHARE: data that the policy explicitly states is not shared/disclosed to third parties.

In the context of our experiment, we merge the subsets into a unified corpus with a single BIO label set so that a single model can jointly learn and distinguish between all four data practice categories. This allows us to assign a fine-grained practice label that captures both first-party collection and third-party disclosure rather than train separate models for each data flow. We then split this merged corpus into training, validation, and test sets to prevent data leakage across respective splits of 70/15/15.

BERT-Based NER Model.

We fine-tune a pretrained BERT model (bert-base-uncased) for token-level classification over the unified BIO label set. Since BERT uses subword tokenization, we first align word-level labels to subword pieces in such a way that word indices can be recovered for each subword. The model is then trained with a token-classification head using cross-entropy loss. We fine-tune for 4 epochs and a small batch size of 8 for training and 16 for evaluation, a weight decay of 0.01, and a learning rate of $3 * 10^{-5}$. At inference time, we return predictions back to the word level by utilizing only the label assigned to the first subword of each token, pre-subword tokenization.

The raw predictions made are then refined using numerous post-processing rules adapted for privacy-policy language. For instance, we use a family-consistency rule that enforces that entity tags within a particular sentence belong either to the COLLECT/NOT_COLLECT family or the SHARE/NOT_SHARE family. Minority family tags are coerced into the dominant family while maintaining polarity (i.e., whether it is affirmative or negative). This is because privacy policies generally do not mention both first-party collection and third-party disclosure within the

same sentence; it is much more common to see patterns such as the mention of an affirmative first-party collection statement and a denial of first-party collection of certain user information in the same sentence. For example, it is rather common to see a statement such as "We collect e-mail addresses from the user, but not physical addresses," in which case "e-mail addresses" would be labeled with the COLLECT tag and "physical addresses" would be labeled with the NOT_COLLECT tag.

In addition, we use negation-aware rules to detect explicit patterns such as "do not share with third parties" or "never disclose to advertisers" to flip COLLECT or SHARE labels to NOT_COLLECT or NOT_SHARE, respectively, when the sentence clearly negates collection or disclosure of user information. We also apply a small gazetteer of terms denoting third-party entities (e.g., "advertisers," "ad networks," "analytics providers") to help further identify SHARE/NOT_SHARE spans when the surrounding context indicates disclosure to an external party. Furthermore, we use a complementary rule to correct cases where "share" verbs are used, but the indirect object associated with the verb is deemed to be the first party. For example, if the phrase "you share your IP address with us" is found in a service's privacy policy, the entity "IP address" should be labeled as COLLECT rather than SHARE, as the user information is still being provided to the first party despite the usage of the phrase "share." Finally, we use auxiliary rules to ensure local BIO tag consistency (i.e., ensuring I-TYPE tags share the same type suffix as the last B-TYPE tag), use a POS tagger to filter out single-token entities that are not nouns, propagate labels across repeated entity phrases within a singular sentence, and attempt to suppress entities in heading and introductory sentences (see Current Limitations and Error Analysis) to further align our system's results with linguistic norms.

Baseline CRF Model.

For comparison, we implement a supervised baseline using a linear-chain Conditional Random Field trained using the same pre-annotated corpus. In this baseline, each token is represented by a feature dictionary that utilizes both general lexical features and privacy-specific indicators. Core features include the original surface form of the token and its lowercase form, capitalization, and prefixes/suffixes found within the token. We also seek to capture local context by considering the previous two and next two tokens (as well as their lowercase forms) as features.

On top of these generic features, we use cues pertaining to sharing, negation, and third-party referencing. For each token, we inspect a ±5-token context window and set binary features that indicate the presence of negation terms, such as "not," "never," or "no." We use a similar ±5-token context window for sharing-related cues, which is activated based on whether any word in the window is part of a hand-crafted set of terms related to sharing or third parties, such as "share," "disclose," or "sell." Additional logic marks explicit n-grams such as "third party" or "third-party." These cues are used to flag specific privacy-related features that exist within the context window. Similarly, we utilize a small hand-made gazetteer of privacy-relevant nouns like "address," "location," and "password" to indicate whether the current token is related to the user's information.

The CRF uses sklearn-crfsuite and is trained using the L-BFGS algorithm using L1 and L2 regularization parameters of 0.1 for both c_1 and c_2 over 100 max iterations and considering all

possible label-to-label transitions. Upon test time, we apply similar family-consistency and override rules as mentioned in the subsection above from our BERT prediction system to the CRF model.

Evaluation and Results.

As mentioned earlier in the Methodology section, we evaluated our models on the merged Pi-Extract corpus using a 70/15/15 train/validation/test split. All systems are trained on the training set, tuned on the validation set, and evaluated on the test set. We measure our system's performance at the token level using metrics of precision, recall, and F1-score. To evaluate category quality, we collapse B-TYPE and I-TYPE labels into a single TYPE label to measure the validity of predicted labels independent of exact span boundaries. Using this method, we compute averaged weighted F1-scores over the four practice categories. In other words, we evaluate three systems: a CRF baseline with handmade features, a raw BERT-based model, and our full BERT-plus-rules hybrid model.

Overall Results.

The performance of our systems is shown in the following table. Table 1 shows token-level precision, recall, and F1-score for each of the four data flow labels, as well as the weighted average over all categories. There are separate columns for the CRF baseline, BERT-only, and BERT-plus-rules hybrid systems. Our hybrid model attains the highest weighted average F1-score of 0.70, which shows improved performance over our baseline model (0.52) and the plain BERT model (0.63). It also shows improvement over Bui et al.'s system, which had an average F1-score of 0.64. Our hybrid model shows consistent gains in F1-score across all labels compared to our other models.

Compared to our CRF baseline model, the BERT-only model shows a substantial improvement in F1-score when marking user information as collected by the first party or shared to third parties, highlighting the advantage of contextual word embeddings over hand-crafted lexical and window-based features with respect to capturing long-range dependencies in the privacy policy domain. With that said, the BERT-only model falls apart when labeling explicit negations. In particular, the BERT-only model has very poor recall performance when labeling datatypes for NOT_COLLECT and NOT_SHARE, as it often predicts affirmative labels when sharing verbs are present or with other similar sentence constructions.

Consequently, we developed our hybrid model with post-processing rules specifically intended to address such issues. As a result, we see large F1 gains on both negative classes without sacrificing F1 performance on positive classes. In fact, when restricting evaluation to only the positive classes (COLLECT and SHARE), macro F1-scores rise from 0.42 to 0.47 for the CRF baseline and from 0.40 to 0.62 for the BERT-only model. In contrast, macro F1 remains unchanged for our hybrid model (0.66 in both cases), indicating that our model performs comparatively well on both affirmative and negative data flow categories instead of heavily relying solely on the easier and more frequent affirmative cases. Regardless, these findings show that while the hybrid model collectively shows strong overall performance, it still lags on SHARE compared to COLLECT and continues to miss some rare negative practices, suggesting that subtle third-party disclosure instances and nuanced negation patterns remain challenging cases for our system.

Table 1: Category-level precision (P), recall (R), and F1-score (F1) for all systems.

	CRF Baseline			BERT Only			BERT + Rules Hybrid		
	P	R	F1	P	R	F1	P	R	F1
COLLECT	0.64	0.53	0.58	0.68	0.76	0.72	0.74	0.72	0.73
NOT_COLLECT	0.68	0.28	0.38	1.00	0.04	0.07	0.64	0.84	0.73
SHARE	0.35	0.35	0.35	0.53	0.54	0.52	0.63	0.54	0.58
NOT_SHARE	0.41	0.30	0.35	0.73	0.17	0.28	0.52	0.71	0.60
Weighted Avg	0.59	0.47	0.52	0.64	0.66	0.63	0.71	0.68	0.70

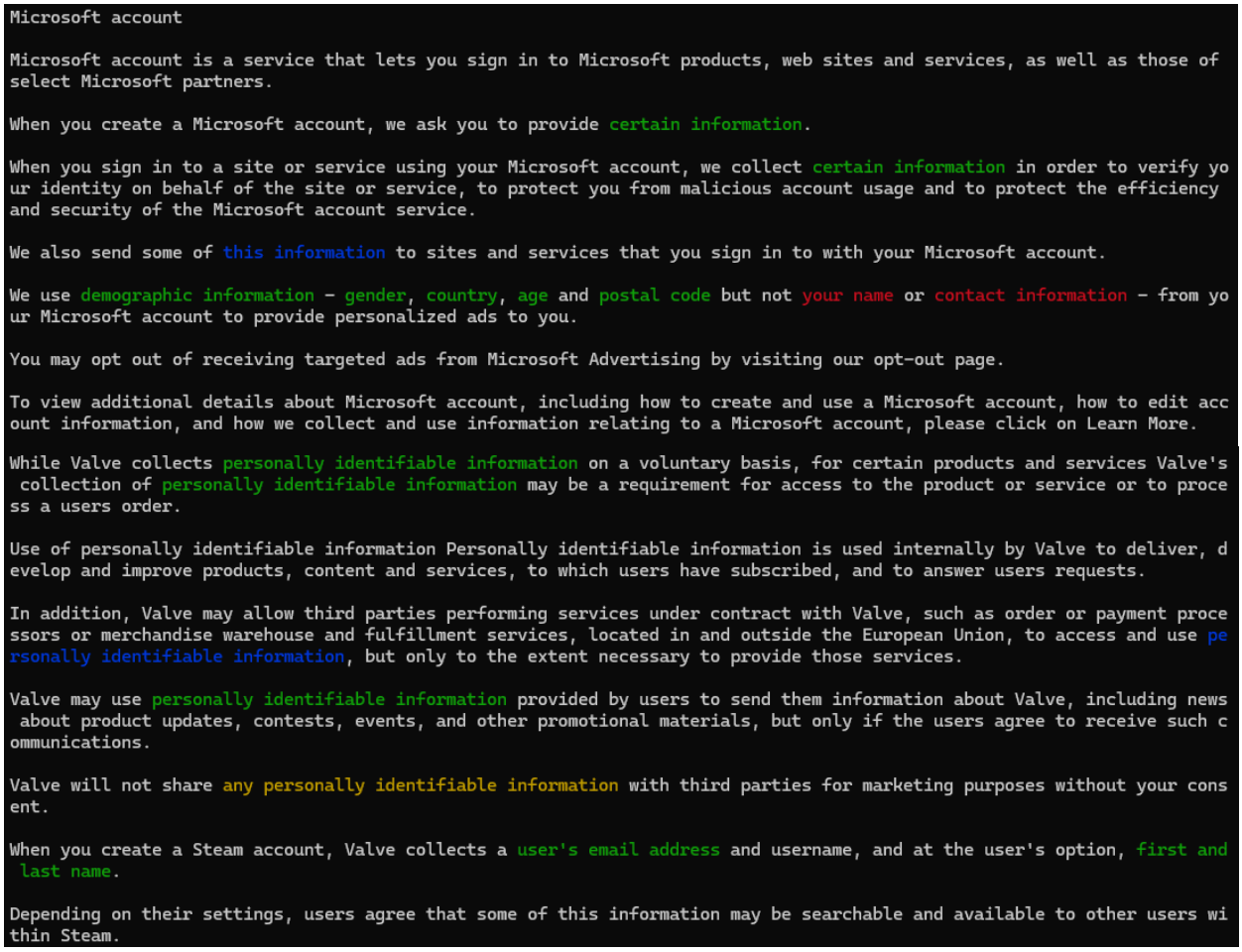

Figure 1: Example system output on privacy policy snippets, where collected (green) and not-collected (red) data items are highlighted for the user. Shared data items (blue) and unshared data items (yellow) are also highlighted for the user.

Error Analysis and Future Work.

Despite the improvements of our hybrid system over our baselines and existing work, there exist limitations with the model. First, performance on SHARE labeling falls behind COLLECT labeling, and the system continues to miss certain rare negative practices. For instance, when third-party disclosure is expressed indirectly or using vague references rather than using explicit sharing verbs. As seen in previous research, it is difficult to effectively capture phrases or sentences in privacy policies that are intentionally vague (Lebanoff and Liu). Sentences such as "Your location information may be used elsewhere on the internet" are inherently unclear and challenging to characterize, even though human understanding may interpret this sentence to mean that one's location information is shared with a third party. Since neither an explicit third party nor a sharing verb is used, our system mislabels "location information" in this sentence for first-party collection. Such errors suggest that subtle third-party disclosure contexts and nuanced negation patterns are still challenging for our system and emphasize the need for richer work on discourse-level modeling in the future.

Second, parts of our post-processing layer rely on hand-crafted rules and gazetteers specifically tailored to English-language privacy policies and tuned to our training and validation corpus. Consequently, our system encodes assumptions about sentence structure and policy style that are subject to change. A limitation inherent to the use of custom-made gazetteers and wordlists is that evolving language will always make such gazetteers non-exhaustive and outdated over time. The same logic applies to our attempts to suppress labeling in introductory statements and headings. Privacy policies frequently use headings and lead-in sentences such as "Information We Collect From the User" or "Cookies are files which may include information such as language preferences," which mention general data practices while only being introductory or explanatory in nature, and therefore should not be labeled. Our model attempts to use formatting-based heuristics to detect these statements, but such cues are style-dependent and are based on our validation set. For example, not every service may format their headings using title case as a capitalization standard like we saw in our validation set, which may result in our model producing false positives. Therefore, we believe that more robust detection of formatting structure and section roles is necessary to distinguish between legitimate data practice mentions and general expository statements.

Third, our sentence-level family-consistency rule assumes that each sentence primarily expresses either first-party collection or third-party disclosure, but not both. In that same vein, our four-way labeling system is inherently flawed in cases where sentences jointly describe multiple actions for the same datatype. For example, the sentence "We collect and share your email address" forces the phrase "email address" to be categorized under a single practice label, despite the information being used for both collection and disclosure purposes. As a result, our model cannot correctly capture multiple simultaneous data flows.

Another limitation with our system arises in the cases of ellipsis and continuity across sentences. Namely, the sentence "You share your personal information with us in various ways." is correctly handled by our post-processing rules to label "personal information" as COLLECT because the indirect object "us" indicates a first-party recipient of user data. This case was originally found in our validation set and our system was tuned accordingly to consider the dependents of a sharing verb for proper classification. However, the immediate next sentence of "For example, you share your location when you log in." will result in "location" being mislabeled under SHARE because the sentence omits the first-party indirect object, and our post-processing rules process sentences in isolation, despite this sentence being semantically

constructed to imply first-party collection according to the previous sentence. Addressing such cross-sentence ellipsis would require moving beyond sentence-level post-processing rules toward discourse-aware post-processing modeling.

Lastly, it is worth noting that our experiments are conducted on a relatively small, expert-annotated corpus, which limits coverage of rarer data flow instances such as those falling under the NOT_SHARE label. As a result, the core BERT model is likely underexposed to the full number of ways denial of sharing can be expressed. Accordingly, since third-party sharing statements (and their negations) are relatively rarer than collection statements, our system scores lower when recognizing and classifying those labels. This necessitates the collection of larger and more diverse annotations of privacy policies to further improve our model's robustness.

Conclusion.

Our work introduces a hybrid NER system for extracting fine-grained data practice categories from privacy policies. We label personal information as collected, not collected, shared, or not shared while differentiating from first- and third-party data flows. Utilizing a BERT token classifier alongside linguistically-informed post-processing rules, our system achieves a token-level weighted F1-score of 0.70 on a modified Pi-Extract corpus, outperforming a CRF baseline and prior neural approaches to the task (Bui et al.).

These results demonstrate that combining contextual word embeddings with negation-aware and dependency-based rules is particularly effective for capturing rare negative practices while maintaining strong performance on affirmative practice classification. In other words, our approach yields a balanced treatment of collection and disclosure statements in both affirmative and negative contexts. With that said, errors related to indirect sharing language, simultaneous data flows, and discourse-level phenomena such as ellipsis highlight examples of challenges for future work. Going forward, extending annotation coverage and incorporating richer discourse and document structure serve as promising foundations for more comprehensive automated analysis of privacy policies.

Author Contributions.

All authors contributed equally to this work. Specific responsibilities include:
Vishram Doodnauth - BERT fine-tuning, corpus merging, subword alignment, proofreading
Samuel Yang - CRF baseline development, paper writing, related work selection
Brian Li - Linguistic post-processing rules, BERT fine-tuning and error analysis, paper writing
Coco Ke - Handcrafted features and wordlists, evaluation metrics design, data pre-processing, proofreading

References.

Brunotte, Wasja, et al. “What about My Privacy? Helping Users Understand Online Privacy Policies.” *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, May 2022, https://doi.org/10.1145/3529320.3529327.

Bui, Duc, et al. “Automated Extraction and Presentation of Data Practices in Privacy Policies.” *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, Jan. 2021, pp. 88–110, https://doi.org/10.2478/popets-2021-0019. Accessed 3 Dec. 2025.

Hosseini, Mitra, et al. “Identifying and Classifying Third-Party Entities in Natural Language Privacy Policies.” *Proceedings of the Second Workshop on Privacy in Natural Language Processing,* Association for Computational Linguistics, 20 Nov. 2020, aclanthology.org/2020.privatenlp-1.3.pdf. Accessed 3 Dec. 2025.

Lebanoff, Logan, and Fei Liu. *Automatic Detection of Vague Words and Sentences in Privacy Policies*. Association for Computational Linguistics, 2018, pp. 3508–17, aclanthology.org/D18-1387.pdf. Accessed 3 Dec. 2025.

Ramanath, Rohan, et al. *Unsupervised Alignment of Privacy Policies Using Hidden Markov Models*. Association for Computational Linguistics, 23 June 2014, pp. 605–10, aclanthology.org/P14-2099.pdf. Accessed 3 Dec. 2025.

Ravichander, Abhilasha, et al. *Question Answering for Privacy Policies: Combining Computational and Legal Perspectives*. 2019, pp. 4947–58, aclanthology.org/D19-1500.pdf. Accessed 3 Dec. 2025.

Šlekytė, Irma. “NordVPN Study Shows: Nine Hours to Read the Privacy Policies of the 20 Most Visited Websites in the US.” NordVPN, 23 Oct. 2023, nordvpn.com/blog/privacy-policy-study-us/.