Automated Detection and Analysis of Data Practices Using A Real-World Corpus

Anonymous ACL submission

Abstract

Privacy policies are crucial for informing users about data practices, yet their length and complexity often deter users from reading them. In this paper, we propose an automated approach to identify and visualize data practices within privacy policies at different levels of detail. Leveraging crowd-sourced annotations from the ToS;DR platform, we experiment with various methods to match policy excerpts with predefined data practice descriptions. We further conduct a case study to evaluate our approach on a real-world policy, demonstrating its effectiveness in simplifying complex policies. Experiments show that our approach accurately matches data practice descriptions with policy excerpts, facilitating the presentation of simplified privacy information to users.

1 Introduction

Do internet users care about their online privacy? While studies have shown that users care about their privacy online (Spiekermann et al., 2001), they are also willing to give away their personal information (Barnes, 2006). Studies have found that this is often because, *privacy policies*, which are intended to inform users about what happens with their data online, are notoriously difficult to understand and time-consuming to read (Obar and Oeldorf-Hirsch, 2018). An average user needs to spend ∼200 hours to read all the policies that they come across each year (McDonald and Cranor, 2008). Moreover, privacy policies are written at a college reading level (Ermakova et al.), use complicated legal jargon and are hard to comprehend

Some studies have attempted to ease the design of privacy policies (Schaub et al., 2015) to make them more comprehensible to users. Kelley et al. (2010) found that presenting privacy information visually in the form of a *privacy nutrition label* significantly improved the ability of users to find and understand privacy information. While companies such as Apple and Google have incorporated a *privacy label* as a way to systematically present users with concise summaries of an app's data practices, they often fail to answer important privacy questions (Zhang and Sadeh, 2023). Moreover, most organizations do not provide access to easy-to-read labels, and scaling the creation of such labels using NLP approaches is a non-trivial task.

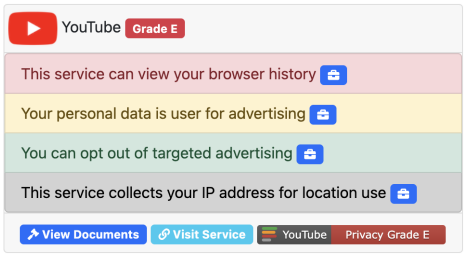

Figure 1: Snapshot of annotated ToS;DR *cases* and assigned privacy letter grade for YouTube

We hence propose to automatically present users information regarding their online privacy using data practice ratings (blocker/good/bad/neutral) and descriptions derived from the Terms of Service; Didn't Read (ToS;DR) platform (Roy et al., 2012). We make the following contributions:

- We manually analyze ToS;DR data descriptions, and cluster similar data practices together[1].
- We create an approach to automatically match excerpts from policies to easily understandable data practice descriptions.
- We design a privacy label to provide users with information regarding their data privacy at various levels of detail.

2 Related Work

Various NLP methods have been employed to analyze privacy policies, yet none offer a holistic comprehension on data practices. Approaches centered around question answering (Ravichander et al.,

[1]Data and code included in the supplementary material

2019; Ahmad et al., 2020), summarization (Zimmeck and Bellovin, 2014; Zaeem et al., 2018) and classification (Wilson et al., 2016; Nokhbeh Zaeem et al., 2022; Tesfay et al., 2018) often lack comprehensive coverage of all data practices outlined in a policy (Nokhbeh Zaeem et al., 2020, 2022). Moreover, these techniques place the onus on users to identify their specific information needs. Users also face the challenge of investing substantial time in deciphering the nuances of each encountered policy (Meier et al., 2020).

Previous studies categorized internet users into 5 categories based on varying levels of privacy concern and online privacy-related behaviors (Kumaraguru and Cranor, 2005; Dupree et al., 2016). Given that each user category necessitates a different level of privacy information, there is a clear need for an approach that addresses this diversity. Our solution caters to this requirement.

3 Methodology

3.1 Dataset

Terms of Service; Didn't read (ToS;DR) aims to make understanding privacy policies easier by crowdsourcing annotations on them (Roy et al., 2012). On the platform, internet users sign up to annotate policies by matching a policy excerpt to a pre-defined data practice (called a "case" on the platform). Once an annotator makes an excerpt-case match, a ToS;DR moderator either accepts the annotation or rejects it with comments. When a threshold percent of moderator approved *case*-excerpt matches are annotated for a policy, it is given a letter grade.

ToS;DR comprises 130 privacy policy-related *cases*. A case on ToS;DR contains an average of 83 policy excerpts (min: 2; max: 83), each from a different service. Eaxt excerpt has an average length of 32.26 words/excerpt. Some *cases* describe similar but contrasting data practices i.e. different approaches to handling the same data type. We deemed these as **contrasting cases**. Following is an example of two contrasting cases:
'*Your personal data is not sold*'
'*Your personal data is sold unless you opt out*'
Two authors manually evaluated the cases in ToS;DR and grouped *contrasting cases* into 24 clusters (min: 2; max: 6 excerpts/cluster). We note that *contrasting cases* might often be a negation of one another, or might differ by describing a special circumstance of a general case, such as,
'*You can delete your content from this service*'
'*You cannot delete your content from this service*'
'*You cannot delete your content, but it makes sense for this service*'
The remaining 67 *cases* had no contrasting data practices. We deemed these as **standalone cases**.

Based on whether *cases* preserve or erode user privacy, they are pre-labeled with one of the following rating categories: *blocker*, *bad*, *neutral*, or *good*. Of the 130 cases, 63 were rated good, 35 bad, 25 neutral, and 7 blockers. Figure 1 shows a snapshot from ToS;DR for the service *YouTube* containing YouTube's assigned letter grade, selected cases, and their respective ratings (Blocker: Red; Bad: Yellow; Good: Green; Neutral: Grey). Ratings indicate respect towards user privacy, where 'blockers' often indicate practices that are aggressively adversarial towards user privacy, while those rated 'good' indicate privacy-preserving practices. 'Bad' and 'neutral' fall in the middle.

3.2 Experiments

In this section, we describe our approach to automatically match excerpts from privacy policies to their respective ToS;DR *cases*. Formally, given an excerpt from a privacy policy, S_j, the task is to find the case(s) C_i that represent the data practice described in the excerpt S_j. Here, C is the set of 130 ToS;DR *cases*; S is the set of excerpts from a privacy policy. This problem can be framed as a multi-label classification task or a binary classification task. Preliminary experiments indicated that the multi-label classification was ineffective, due to a substantial imbalance in class distribution. Therefore, choose a binary classification technique to train a model to distinguish between matching and disparate case-excerpt pairs.

Annotated case-excerpt pairs from ToS;DR serve as positive training samples. To collect negative samples, we use two sampling approaches, random sampling (RS) and cluster-based sampling (CBS).
Random Sampling (RS): For each annotated ToS;DR *case*-excerpt pair, we randomly sample excerpts matched with other *case* to serve as a negative sample irrespective of it being a *contrasting* or *standalone case*.
Cluster-Based Sampling (CBS): For each annotated *case*-excerpt pair, if the *case* is part of a cluster with contrasting *cases*, we first sample excerpts matched with other cases within the same cluster. We do this until we have either exhausted all pos-

Model	C	Standalone Set				Contrasting Set			
		P	R	F1	S	P	R	F1	S
RS	0	**0.94**	0.95	**0.94**	2091	**0.75**	0.37	0.50	557
	1	0.95	**0.94**	**0.94**	2091	0.72	0.93	0.81	996
CBS	0	0.89	**0.95**	0.92	2091	0.71	**0.82**	**0.76**	557
	1	**0.95**	0.88	0.91	2091	**0.95**	**0.91**	**0.93**	996

Table 1: Performance of PrivBERT on random sampling and cluster-based sampling with a 1:3 sampling ratio (P: Precision, R: Recall, S: Support)

sible unique negative samples[2] or we have the required number of negative samples. When the number of negative samples is not satisfied, we sample from outside the cluster randomly. For standalone cases, we collect negative samples randomly

We trained separate models for each sampling technique and sampling ratio. We experiment with sampling 1x, 2x, 3x, and 5x number of negative samples as positive samples. We divided the data into train, validation, and test sets in the ratio 3:1:1. We keep the test and validation sets constant across different sampling ratios, with 1:1 sampling. To train the model, we input the concatenated case C_i and segment S_j separated by a special token to fine-tune PrivBERT, a privacy policy language model (Srinath et al., 2021). We use a binary cross-entropy loss function to optimize the parameters and train the model using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-5 for 3 epochs.

We evaluate the trained model on test sets each containing policy excerpts previously unseen by the model. The standalone set and the contrasting set contain case-excerpt pairs from standalone cases contrasting cases respectively. We compare the performance of our model to several approaches listed in Table 2. 1) We use Sentence-BERT (Reimers and Gurevych, 2019) to vectorize policy excerpts and ToS;DR cases. We then calculate the cosine distance between each pair and identify a threshold similarity score (0.25) based on the validation set. 2) We prompt OpenAI's GPT-4 to identify matching case-excerpt pairs 3) We enhance the GPT-4 prompt with two examples, one positive/negative case-excerpt pair 4) We train RoBERTa (Liu et al., 2019) recreating the experiments with PrivBERT and report the results on the best performing model.

4 Results

The results of PrivBERT trained on datasets created by random sampling (RS) and cluster-based sampling (CBS), for a 1:3 sampling ratio, is shown in Table 4. The results in the table are reported after taking the mean (max standard deviation: 0.05) on 3 runs with different random initial states and training sets shuffled. The 0 category refers to disparate case-excerpt pairs while 1 refers to matching pairs. The table shows that the results on the standalone set are superior to those on the contrasting case set across both sampling techniques. This is likely since *contrasting cases* are often lexically and semantically quite similar, and therefore a harder problem. We also see that cluster-based sampling performs significantly better than random sampling on the contrasting set while maintaining a similar performance on the standalone set. This is expected since a larger number of samples are used to train the model to solve the harder problem of disambiguating between similar cases.

For the RS model, the results on the contrasting set improved with larger negative sampling ratios while the CBS model remained the same. This could indicate that a larger number of contrasting case training samples were necessary for the RS model. On the other hand, the results for both the models on the standalone set remained constant across different sampling ratios. Results for different sampling ratios are presented in the appendix.

Model	C	P	R	F1
Sentence-BERT	0	0.61	0.66	0.63
	1	0.78	0.71	0.74
GPT-4 (Zero-shot)	0	0.72	0.74	0.73
	1	0.82	0.80	0.81
GPT-4 (2-shot)	0	0.79	0.74	0.76
	1	0.83	0.86	0.84
RoBERTa (CS)	0	0.78	0.85	0.81
	1	0.91	0.87	0.89
PrivBERT (CS)	0	**0.80**	**0.88**	**0.83**
	1	**0.95**	**0.89**	**0.92**

Table 2: Test performance (P: Precision, R: Recall)

We achieve state-of-the-art results on this dataset, the performance of several other approaches is

[2]This can happen when only one of the *cases* in a cluster has a large number of positive samples

shown in Table 2. Finetuned RoBERTa-base is the second best-performing model, followed by the two-shot version of GPT-4. The performance improvement due to fine-tuning depicts the complexity of the task. Sentence-BERT performs well on case-excerpt pairs with significant lexical overlap.

Under real-world conditions, identifying all data practices within a privacy policy would involve testing whether an excerpt matches any case in the full set of ToS;DR cases. It is therefore important that the model be trained robustly on dissimilar pairs, supporting training on larger sampling ratios and providing high precision scores for negatives. Furthermore, the erroneous classification of a pair as a 'match' can be readily discerned and flagged, thus facilitating the establishment of a feedback loop. Conversely, the misclassification of a pair as dissimilar may evade detection, reinforcing our argument for the necessity of robustly performing models on dissimilar pairs.

4.1 Case Study

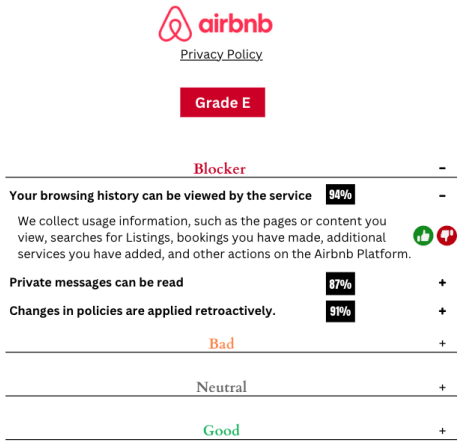

Figure 2: Privacy label with varying information levels. Detailed information access by the expand/contract feature. Model probabilities shown as %; thumbs-up/down icons for user feedback on matches.

We present a case study evaluating our model's performance on a previously un-annotated policy. We chose Airbnb[3], a popular online marketplace, due to its popularity and its legitimate need to access sensitive personal user information. We split the policy based on new line characters, resulting in 150 segments encompassing paragraphs, section titles, and list items. Two authors of the paper thoroughly reviewed each segment and tagged them with any associated ToS;DR *cases*. Subsequently, we employed the best-performing CBS model to identify matching *case*-excerpt pairs. Table 3 that the model consistently identified true positives, while producing a few false positives, therefore leading to a high recall for matching case-excerpt pairs. Most false positives were caused due to inaccurate predictions on privacy policy section titles. The model was able to identify true negatives with a high degree of accuracy while producing a few false negatives, leading to high precision scores for dissimilar case-excerpt pairs.

	Precision	**Recall**	**F1**	**Support**
True	0.79	0.88	0.83	285
False	0.87	0.84	0.85	19,215

Table 3: Case-level results

In Fig. 2, we demonstrate the tool that provides users with privacy information at varying levels of granularity. At the highest level, identified *cases* are summarized using a letter grade with the grading scheme adopted by ToS;DR: 1 or more 'blockers': Grade E || > 75% 'bad' *cases*: Grade D || 50% - 75% 'bad' cases: Grade C || < 50% 'bad' cases: Grade B || < 25% 'bad' cases: Grade A. At the next level, identified *cases* are listed under their rating categories. Finally, at the most fine-grained level policy excerpts associated with a case are listed along with a match probability score.

5 Conclusion

Our research introduces a novel solution to the challenge of simplifying complex privacy policies. We provide an automated approach to scale identification of data practices in privacy policies, providing users with information regarding their data privacy. By automating the process of aligning data practices with policy excerpts, we are taking a crucial step towards eliminating the need for users to laboriously parse through lengthy and intricate documents, ensuring that they receive precise and easily digestible information about their online privacy. This lowers the barrier for internet users to understand what is happening with their data online, thereby allowing them to make informed privacy decisions. While there is room for further refinements and research, our model presents a promising foundation for future advancements in the field of usable privacy, ultimately promoting greater user empowerment and security in the digital age.

[3]https://www.airbnb.com/

6 Limitations

Here, we attempt to match ToS;DR cases to their corresponding privacy policy segments. While the dataset we create attempts to simulate real-world conditions through the hold-out set, we match segments that are preciously identified by annotators. It is therefore possible that there is some loss of performance in a real-world setting. We rectify this issue in our case study by applying our model on automatically segmented policies and find that the performance is still comparable with that of the hold-out set. However, it is possible that the results in the case study might not scale over all privacy policies.

We use the rating scheme developed by ToS;DR which could potentially be biased. There is therefore a need for further research to understand user privacy preferences towards threshold for blocker, bad, good and neutral data practices. This extends to the assignment of letter grades.

7 Ethical Statement

We do not foresee any ethical issues with our work.

References

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. Policyqa: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749.

Susan B Barnes. 2006. A privacy paradox: Social networking in the united states. *First Monday*, 11(9).

Janna Lynn Dupree, Richard Devries, Daniel M Berry, and Edward Lank. 2016. Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5228–5239.

Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. Readability of privacy policies of healthcare websites.

Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 1573–1582. ACM.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ponnurangam Kumaraguru and Lorrie Faith Cranor. 2005. Privacy indexes: a survey of westin's studies.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp*, 4:543.

Yannic Meier, Johanna Schäwel, and Nicole C Krämer. 2020. The shorter the better? effects of privacy policy length on online privacy decision-making. *Media and Communication*, 8(2):291–301.

Razieh Nokhbeh Zaeem, Ahmad Ahbab, Josh Bestor, Hussam H Djadi, Sunny Kharel, Victor Lai, Nick Wang, and K Suzanne Barber. 2022. Privacycheck v3: Empowering users with higher-level understanding of privacy policies. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1593–1596.

Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K Suzanne Barber. 2020. Privacycheck v2: A tool that recaps privacy policies for you. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3441–3444.

Jonathan A Obar and Anne Oeldorf-Hirsch. 2018. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, pages 1–20.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. 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 4949–4959.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Hugo Roy, JC Borchardt, I McGowan, J Stout, and S Azmayesh. 2012. Terms of service; didn't read. *Web Page, June. URL https://tosdr. org*.

Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*, pages 1–17.

Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. 2001. E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 38–47.

Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at scale: Introducing the privaseer corpus of web privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6829–6839.

Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. Privacyguide: towards an implementation of the eu gdpr on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 15–21.

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, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):1–18.

Shikun Zhang and Norman Sadeh. 2023. Do privacy labels answer users' privacy questions. In *Network and Distributed System Security Symposium*.

Sebastian Zimmeck and Steven M Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 1–16.

A Appendix

A.1 Loss Function

Given a *case* from ToS;DR, C_i, and an excerpt from a privacy policy, S_j, where,

$C \in$ set of ToS;DR *cases*;
$S \in$ set of privacy policy excerpts

The loss function is given by,

$$-y \log(\hat{y}) + (1-y) \log(1-\hat{y}) \quad (1)$$

$$y = Match(C_i, S_j) \quad (2)$$

A.2 Privacy Rating Analysis

We further investigated each ToS;DR privacy category, blocker, bad, good, and neutral by calculating the perplexity of the excerpts associated with them. Figure 3 shows the average perplexity of excerpts calculated using RoBERTa, a general purpose language model and PrivBERT, a language model pretrained on privacy policy text. For comparison, we randomly sampled 1000 sentences from ten most visited websites' policies. We calculated the perplexity by masking each word in the sequence separately and exponentiating the average the negative log likelihood of size 8 sequences with a stride of 4.

From Figure 3, it is evident that PrivBERT exhibits notably reduced perplexity scores in comparison to RoBERTa, a phenomenon attributed to PrivBERT's pretraining on policy text. We hypothesize that for PrivBERT (but not RoBERTa), text segments commonly encountered in privacy policies should yield lower perplexity scores, while those that are infrequent should result in higher perplexity scores. Figure 3 shows that 'blockers' have the highest sentence perplexity scores for both language models suggesting that they might be written in particularly convoluted language unusual even within privacy policies. Conversely, other categories exhibit closely aligned perplexity scores for RoBERTa, but distinct variations for PrivBERT, suggesting that the pretraining data may encompass text associated with each category at varying frequencies. Notably, 'bad' practices are the second most perplexing category for RoBERTa, yet the least perplexing for PrivBERT, signifying the frequent inclusion of 'bad' practices-related text in policies. Conversely, this trend is reversed for 'good' practices, indicating their relative rarity.

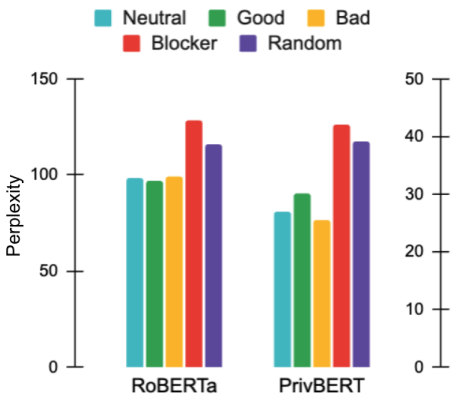

Figure 3: Perplexity distribution of ToS;DR ratings

Ratio	Model	C	Standalone Set				Contrasting Set			
			P	R	F1	S	P	R	F1	S
1:1	RS	0	0.94	0.93	0.93	2091	0.72	0.28	0.40	557
		1	0.93	0.94	0.94	2091	0.70	0.93	0.80	996
	CBS	0	0.93	0.93	0.93	2091	0.77	0.60	0.68	557
		1	0.93	0.92	0.92	2091	0.90	0.95	0.92	996
1:2	RS	0	0.95	0.95	0.95	2091	0.75	0.30	0.43	557
		1	0.95	0.95	0.95	2091	0.70	0.94	0.81	996
	CBS	0	0.87	0.94	0.90	2091	0.70	0.81	0.75	557
		1	0.94	0.86	0.90	2091	0.94	0.91	0.92	996
1:3	RS	0	0.94	0.95	0.94	2091	0.75	0.37	0.50	557
		1	0.95	0.94	0.94	2091	0.72	0.93	0.81	996
	CBS	0	0.89	0.95	0.92	2091	0.71	0.82	0.76	557
		1	0.95	0.88	0.91	2091	0.95	0.91	0.93	996
1:5	RS	0	0.95	0.95	0.95	2091	0.69	0.45	0.54	557
		1	0.93	0.94	0.93	2091	0.70	0.92	0.78	996
	CBS	0	0.86	0.96	0.90	2091	0.71	0.83	0.76	557
		1	0.92	0.88	0.89	2091	0.90	0.90	0.90	996

Table 4: Performance of PrivBERT on random sampling and cluster-based sampling over various sampling ratios (P: Precision, R: Recall, S: Support)

A.3 Prompts Used for GPT-4

We used the OpenAI API GPT-4 models with a temperature setting of 0.

You are a privacy policy expert. Given a title and quote, your task is to evaluate whether the title represents the data practice described in the quote. Your output should only be either 0 (indicating the title does not represent the quote) or 1 (indicating the title represents the quote).

Example 1
Title: Third parties are involved in operating the service.
Quote: Note that we don't use any 3rd party website statistics tools like Google Analytics or similar.
Output: 1

Example 2
Title: You can opt out of targeted advertising
Quote: Our CDN is Cloudflare, and they may include cookies with our pages to provide a better service.
Output: 0

Title: <title>
Quote: <quote>