Retrieval Enhanced Data Augmentation for Question Answering on Privacy Policies

Anonymous ACL submission

Abstract

Prior studies in privacy policies frame the question answering (QA) tasks as identifying the most relevant text segment or a list of sentences from the policy document for a user query. However, annotating such a dataset is challenging as it requires specific domain expertise (e.g., law academics). Even if we manage a small-scale one, a bottleneck that remains is that the labeled data are heavily imbalanced (only a few segments are relevant) –limiting the gain in this domain. Therefore, in this paper, we develop a novel data augmentation framework based on ensembling retriever models that captures the relevant text segments from unlabeled policy documents and expand the positive examples in the training set. In addition, to improve the diversity and quality of the augmented data, we leverage multiple pre-trained language models (LMs) and cascaded them with noise reduction oracles. Using our augmented data on the PrivacyQA benchmark, we elevate the existing baseline by a large margin (11% F1) and achieve a new state-of-the-art F1 score of 50%. Our ablation studies provide further insights into the effectiveness of our approach.

Segmented policy document $\boldsymbol{S}$
($\boldsymbol{s}_1$) We do not sell or rent your personal information to third parties for their direct marketing purposes without your explicit consent. ($\boldsymbol{s}_n$) ...We will not let any other person, including sellers and buyers, contact you, other than through your ...
Queries $\boldsymbol{I}$ annotating the red segment as irrelevant
($\boldsymbol{i}_1$) How does Fiverr protect freelancers' personal information? ($\boldsymbol{i}_2$) What type of identifiable information is passed between users on the platform?
Queries $\boldsymbol{R}$ annotating the red segment as relevant
($\boldsymbol{r}_1$) What are the app's permissions? ($\boldsymbol{r}_2$) What type of permissions does the app require?
Queries $\boldsymbol{D}$ that annotators disagree about relevance
($\boldsymbol{d}_1$) Do you sell my information to third parties? ($\boldsymbol{d}_2$) is my information sold to any third parties?

Table 1: QA (sentence selection) from a policy document $\boldsymbol{S}$. **Sensitive**: For queries $\boldsymbol{R}$ and $\boldsymbol{I}$, annotators at large tagged sentence $\boldsymbol{s}_1$ as relevant, and irrelevant respectively. On the other hand, sentence $\boldsymbol{s}_n$, though analogous to $\boldsymbol{s}_1$ in meaning, was never tagged as relevant. **Ambiguous**: For queries $\boldsymbol{D}$, experts interpret $\boldsymbol{s}_1$ differently and disagree on their annotations.

1 Introduction

Understanding privacy policies that describe how user data is collected, managed, and used by the respective service providers is crucial for determining if the conditions outlined are acceptable. Policy documents, however, are lengthy, verbose, equivocal, hard to understand (McDonald and Cranor, 2008; Reidenberg et al., 2016). Consequently, they are often ignored and skipped by users (Commission et al., 2012; Gluck et al., 2016).

To help the users better understand their rights, privacy policy QAs are framed as answer sentence selection tasks, essentially a binary classification task to identify if a policy text segment is relevant or not (Harkous et al., 2018). However, annotating policy documents requires expertise and domain knowledge, and hence, it is costly and hard to obtain. Moreover, as most texts in policy documents are not relevant, the data is heavily imbalanced. For example, the only existing dataset, PrivacyQA (Ravichander et al., 2019) has 1,350 questions in the training dataset, and the average number of answer sentences is 5, while the average length of policy documents is 138 sentences.

In this work, we mitigate data imbalance by augmenting positive QA examples in the training set. Specifically, we develop automatic retrieval models to supplement retrieved relevant policy sentences for each user query. The queries we keep unchanged as they are often less variant and limited to a few forms (Wilson et al., 2016).

Augmenting privacy policies is challenging. First, privacy statements often describe similar information (Hosseini et al., 2016). Thus, their annotations are sensitive to small changes in the text (see Table 1), which may not be tackled using the existing augmentation methods based on data synthesis. For example, Kumar et al. (2020) identifies that even linguistically coherent instances augmented via generative models such as GPT-2 (Anaby-Tavor et al., 2020) do not preserve the class labels well. Hence, we consider a retrieval-based approach to augment the real policy statements to address this. Given a pre-trained LM and a small QA dataset, we first build a dense sentence retriever (Karpukhin et al., 2020). Next, leveraging an unlabeled policy corpus with 0.6M sentences crawled from web applications, we perform a coarse one-shot sentence retrieval for each query in the QA training set. To filter the noisy candidates retrieved, we then train a QA model (as an oracle) using the same pre-trained LM and data and couple it with the retriever.

Second, privacy policies are ambiguous; even skilled annotators dispute their interpretations, e.g., for at least 26% questions in PrivacyQA, experts disagree on their annotations (see Table 1). Therefore, a single retriever model may not capture all relevant policy segments. To combat this, we propose a novel retriever ensemble technique. Different pre-trained models learn distinct language representations due to their pre-training objectives, and hence, retriever models built on them can retrieve a disjoint set of candidates (verified in Section 3). Therefore, we build our retrievers and oracles based on multiple different pre-trained LMs (See Figure 1). Finally, we train a user-defined QA model on the aggregated corpus using them.

We evaluate our framework on the PrivacyQA benchmark. We elevate the state-of-the-art performance significantly (11% F1) and achieve a new one (50% F1). Furthermore, our ablation studies provide an insightful understanding of our model. All data/code will be released upon acceptance.

2 Methodology

The privacy policy QA is a binary classification task that takes a user query q, a sentence p from policy documents and output a binary label $z \in \{0, 1\}$ that indicates if q and p are relevant or not. As most sentences p are labelled as negative, our goal is to retrieve relevant sentences to augment the training data and mitigate the data imbalance issue. Given a QA training dataset $D = \{(q_i, p_i, z_i)\}_{i=1}^{m}$, for each question in D, we (1) retrieve positive sentences from a large unlabeled corpus. (2) filter the noisy

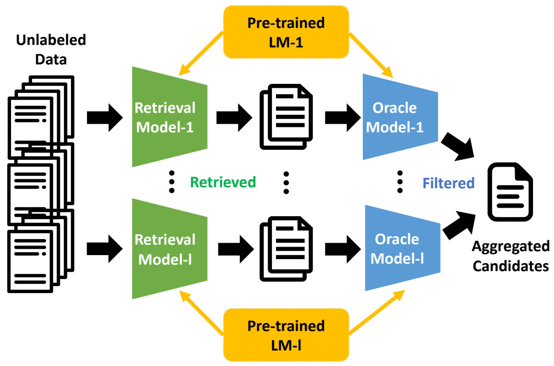

Figure 1: Our framework. Given a pre-trained LM, we train a (i) retriever, (ii) QA model (oracle) both on the small-size labeled data. From an unlabeled corpus, we first, retrieve the coarse relevant sentences (positive examples) for the queries in the training set and use the oracle to filter out noisy ones. We repeat this for multiple different pre-trained LMs. Finally we aggregate them to expand the positive examples in the training set and learn any user-defined final QA model.

examples using oracle models and aggregate final candidates. The final candidates are combined with the base data D to train the QA models. We use an ensemble of retrievers and oracles built upon various pre-trained LMs throughout the whole process. Next, we provide more details of our approach.

Retriever. Our retriever module is built upon the Dense Passage Retriever (DPR) model (Karpukhin et al., 2020; Parvez et al., 2021). It consists of two encoders $Q(\cdot)$ and $P(\cdot)$ that encode the queries and the policy sentences, respectively. The relevance of a query q and a policy sentence p is calculated by the dot product of $Q(q)$ and $P(p)$, i.e., $\text{sim}(q, p) = Q(q)^T \cdot P(p)$. For each positive pair in D, DPR optimizes the cross-entropy loss with in-batch negatives (Henderson et al., 2017). We train a retriever $\mathcal{R}_L$ on D, where the encoders in $\mathcal{R}_L$ are initialized with a pre-trained LM L. At inference, $\mathcal{R}_L$ retrieves the top-k most relevant policy sentences from an unlabeled corpus of policy sentences $\mathcal{P} = \{p_1, \ldots, p_M\}$ for each query q_i in D, i.e., $\mathcal{R}_L(\{q_i\}_{i=1}^{m}, \mathcal{P}, k) = \{(q_i, p_j, 1) : i \in [m], p_j \in \mathcal{P}_{\text{top}}(q_i, k)\}$, where $\mathcal{P}_{\text{top}}(q_i, k) := \arg\max_{\mathcal{P}' \subset \mathcal{P}, |\mathcal{P}'|=k} \sum_{p \in \mathcal{P}'} \text{sim}(q_i, p)$.

Filtering Oracle. To filter out the noisy retrievals from $\mathcal{R}_L(\{q_i\}_{i=1}^{m}, \mathcal{P}, k)$, we train a QA model ($\mathcal{Q}_L$) using the training data D as an oracle to predict whether a query q and a (retrieved) policy sentence p are relevant or not (i.e., $\mathcal{Q}_L(q, p) \in \{0, 1\}$). Note that both $\mathcal{Q}_L$ and retriever $\mathcal{R}_L$ are built upon the same pre-trained LM L but differ in training objectives (e.g., ranking problem vs. binary classification and, w/ and w/o in-batch negatives) and model architectures. We verify the effectiveness of oracle filtering in Section 3.2. We denote retrieval outputs after filtering as $\mathcal{D}_L = \{(q, p, 1) : \mathcal{Q}_L(q, p) = 1, \forall (q, p, 1) \in \mathcal{R}_L(\{q_i\}_{i=1}^{m}, \mathcal{P}, k)\}$.

Ensemble. Unlike other NLP domains, a privacy policy sentence can frequently have multiple interpretations (see Table 1). Hence, a single retrieved corpus $\mathcal{D}_L$ may not capture all relevant candidates covering such diverse interpretations. To this end, we use a set of pre-trained LMs $\mathcal{L} = \{L_1, \ldots, L_l\}$ and aggregate all the corresponding retrieved corpora, $\mathcal{D}_{\text{aug}} = \bigcup_{L \in \mathcal{L}} \mathcal{D}_L$. In Section 3, we show that retrieved corpora using multiple pre-trained LMs with different learning objectives can bring a different set of relevant candidates. Lastly, we aggregate $\mathcal{D}_{\text{aug}}$ with D (i.e., final train corpus $\mathcal{T} = \mathcal{D}_{\text{aug}} \cup D$) and train our final QA model with user specifications (e.g., architecture, pre-trained LM).

3 Experiments

In this section, we evaluate our approach and present the findings from our analysis.

Settings. We evaluate our approach on PrivacyQA benchmark and recall that this is in fact a text classification task. Following Ravichander et al. (2019), we use *precision*, *recall*, and *F1 score* as the evaluation metrics. As for the retrieval database $\mathcal{P}$, we crawl privacy policies from the most popular mobile apps spanning different app categories in the Google Play Store and end up with 6544 documents (0.6M statements). By default, all retrievals use top-10 candidates w/o filtering. All data/models/codes are implemented using (i) Huggingface Transformers (Wolf et al., 2019), (ii) DPR (Karpukhin et al., 2020) libraries and will be released.

Baselines. We fine-tune three pre-trained LMs on PrivacyQA as baselines: (i) *BERT*: Our first baseline is BERT-base-uncased (Devlin et al., 2019) which is pre-trained on generic NLP textual data. (ii) *PBERT*: We adapt *BERT* to the privacy domain by fine-tuning it using masked language modeling on a corpus of 130k privacy policies (137M words) collected from apps in the Google Play Store (Harkous et al., 2018). (iii) *SimCSE*: We take the *PrivacyBERT* model and apply the unsupervised contrasting learning SimCSE (Gao et al., 2021) model on the same 130k privacy policy corpus.

We also consider three other retrieval augmented QA models based on individual pre-trained LM without ensemble: (iv) *BERT-R*: $\mathcal{L} = \{BERT\}$, (v) *PBERT-R*: $\mathcal{L} = \{PBERT\}$, (vi) *SimCSE-R*: $\mathcal{L} = \{SimCSE\}$. We first construct $\mathcal{T}$ (both settings: w/ and w/o oracle) and fine-tune on it the corresponding pre-trained LM as the final QA model. Finally, we consider one more ensemble retrieval augmented baseline (vii) *Baseline-E*, which is exactly the same as ours (settings below) except there are no intermediate filtering oracles.

Method	Oracle	Precision	Recall	F1
Human	-	68.8	69.0	68.9
W/o data augmentation				
BERT + Unans.	-	44.3	36.1	39.8
BERT (reprod.)	-	47.5	38.5	42.5
PBERT	-	50.8	43.1	46.7
SimCSE	-	48.8	42.2	45.3
Retriever augmented				
BERT-R	✗	39.9	50.8	44.7
	✓	46.5	45.5	46.0
PBERT-R	✗	48.4	45.6	46.9
	✓	49.5	46.3	47.8
SimCSE-R	✗	48.4	47.2	47.8
	✓	51.0	45.2	47.9
Ensemble retriever augmented				
Baseline-E	✗	23.0	54.3	32.3
ERA	✓	47.1	52.9	**49.8**
ERA-D	✓	51.3	50.0	**50.6**

Table 2: Test performances on PrivacyQA. BERT + Unans. refers to Ravichander et al. (2019). Retrieved candidates improves all the baseline QA models, especially when being filtered. Our ensemble retriever approach combines them and achieve the highest gains.

Ours. We construct out augmented corpus $\mathcal{T}$, discussed in Section 2, using the (i) all three aforementioned pre-trained LMs: $\mathcal{L} = \{BERT, PBERT, SimCSE\}$ (ii) domain adapted models only: $\mathcal{L} = \{PBERT, SimCSE\}$. For brevity, we call them: Ensemble Retriever Aug. (ERA) and ERA-D. By default, we fine-tune *SimCSE* as the final QA model.

3.1 Results and Analysis

The results are listed in Table 2. Overall, domain adapted models *PBERT* and *SimCSE* excel better than the generic *BERT* model. The retrieval augmented models enhance the performances more, specially the recall score. However, they might contain several noisy examples (see Table 4), and filtering those out improves the precision scores. Finally, ERA and ERA-D aggregate these high-quality filtered policies–leading toward the highest gain (11% F1 from the previous baseline) and a new state-of-the-art result with an F1 score of 50.6.

Table 3 shows the performance breakdown for

Query Type	%	B	PB	S	ERA
Data Collection	42	45	**46**	**46**	**48**
Data Sharing	25	**43**	37	41	**43**
Data Security	11	**65**	61	60	60
Data Retention	4	**52**	35	35	**56**
User Access	2	**72**	48	31	61
User Choice	7	41	**60**	42	31
Others	9	36	45	**52**	**55**
Overall	100	45	47	48	**50**

Table 3: Breakdown of F1-score. B, PB, S refers to retrievers *BERT-R*, *PBERT-R*, and *SimCSE-R*. Different models performs better for different types (black-bold). Our framework ERA combines them and enhances performances for all categories, in general (except: red).

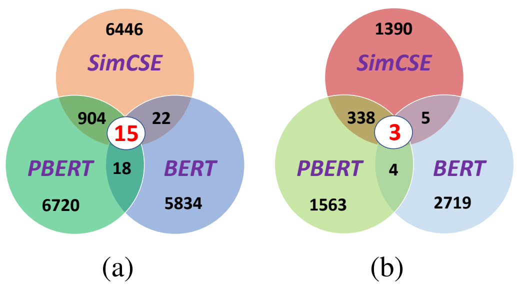

Figure 2: Venn diagram of low mutual agreement (<1%) among retrievers (a); even amplified after filtering (b).

different query types. Individual retrieval augmented models perform at different scales for each type, and combining their expertise, ERA enhances the performances for all types. Next, we show the Venn diagram of overlapping retrievals in Figure 2. Although being retrieved from the same corpus, they rarely overlap. This validates our hypothesis that retrievers built upon different pre-trained LMs learn distinct language representations and retrieve diverse candidates.

3.2 Ablation Study

Are oracles needed? From Table 3, in general, aggregating retrievals with oracle filtering enhances model performances than crude additions.

A common oracle. Performances of ERA (last row in Table 3) with a common oracle based on *SimCSE* for all the retrievers regardless of their corresponding pre-trained models are 49.2, 45.2, and 47.1, respectively–validating the requirement of filtering using the corresponding pre-trained LM.

Other pre-trained LM as the final QA model. Fine-tuning *PBERT* instead of *SimCSE* on $\mathcal{T}$ (last two rows in Table 2) becomes: 47.0, 47.1, 47.0 and 51.0, 45.9, 48.3, respectively. This shows that our approach is generic and enhances the performance regardless of the end model.

Q: who all has access to my medical information?

Correct Retrievals: (S) We may share your information with other health care providers, laboratories, govt. agencies, insurance companies, organ procurement organizations, or medical examiners. (P) Lab, Inc will transmit personal health information to authorized medical providers.

Incorrect Retrievals: (S) However, we take the protection of your private health information very seriously. (P) All doctors, and many other healthcare professionals, are included in our database.

Table 4: Retrieval examples: S (*SimCSE-R*), P (*PBERT-R*).

Which pre-trained LMs to use? Table 3 shows ERA-D that combines fewer number of pre-trained LMs may even outperform the one with more models, ERA. Though here we consider a simple approach (in-domain) for selecting the potential subset of models, this paves a new direction of future research (e.g., Parvez and Chang (2021)).

Qualitative examples. Table 4 (more in Appendix) shows some example retrievals of different models. They are distinct from expert annotated ones and can bring auxiliary knowledge.

4 Related Works

A line of works focuses on using NLP techniques for privacy policies (Wilson et al., 2016; Harkous et al., 2018; Zimmeck et al., 2019; Bui et al., 2021; Ahmad et al., 2021). Besides the QA tasks as sentence selection, Ahmad et al. (2020) propose another SQuAD-like (Rajpurkar et al., 2016) privacy policy reading comprehension dataset for a limited number of queries. Oppositely, we focus on the more challenging one, which allows unanswerable questions and "non-contiguous" answer (Ravichander et al., 2021). Model or data aggregation has also been studied under different NLP contexts (e.g., bagging (Breiman, 1996), meta learning (Parvez et al., 2019)). Here, we aggregate the retriever outputs using different pre-trained LMs.

5 Conclusion

We develop a noise-reduced retrieval-based data augmentation method that uses the combination of different pre-trained language models as a backbone. Although we focus on the privacy policy domain, our approach is generic and can broadly be applied to other NLP domains. We will leave the exploration as future work.

References

Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for 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 4402–4417.

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.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110.

Federal Trade Commission et al. 2012. Protecting consumer privacy in an era of rapid change. *FTC report*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Mitra Bokaei Hosseini, Sudarshan Wadkar, Travis D Breaux, and Jianwei Niu. 2016. Lexical similarity of information type hypernyms, meronyms and synonyms in privacy policies. In *2016 AAAI Fall Symposium Series*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

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

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md Rizwan Parvez, Tolga Bolukbasi, Kai-Wei Chang, and Venkatesh Saligrama. 2019. Robust text classifier on test-time budgets. 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 1167–1172.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. 2021. Breaking down walls of text: How can NLP benefit consumer privacy? 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 4125–4140, Online. Association for Computational Linguistics.

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 4947–4958.

Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. 2016. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190.

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. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86.

Supplementary Material: Appendices

A Limitations/Reproduction

In this paper, we show that leveraging multiple different pre-trained LMs can augment high-quality training examples and enhance the QA (sentence selection) task on privacy policies. Our approach is generic and such unification of different kinds of pre-trained language models for text data augmentation can improve many other low-resourced tasks or domains. However, it is possible that our approach:

- may not work well on other scenarios (e.g., domains/language or tasks etc.,).
- subject to the choice of particular set of models. For example, as mentioned in Section, 3.2, fine-tuning pre-trained models other than *SimCSE* (Gao et al., 2021) as the final QA model achieve lower gain.
- may not work for certain top-k retrievals. For example, from Table 6, we get different results with different scales for variable top-k values (e.g., top-10, top-100).
- uses the same set of hyperparameters for all:
  - QA model:
    - learning rate: $2e^{-5}$,
    - train epoch: 4,
    - per gpu train batch size: 31,
    - num gpus: 4
    - fp16 enabled
    - others: mostly default as in Huggingface
    - train time: around 2 hours
    - Higgingface transformer version 0.3.2. (it has Apache License 2.0)
  - Retriever model:
    - learning rate: $2e^{-5}$,
    - train batch size: 16,
    - train epoch: 100,
    - global_loss_buf_sz 600000,
    - others: mostly default as in DPR (It has Attribution-NonCommercial 4.0 International license)
    - num gpus: 3
    - Higgingface transformer version 0.3.2 (it has Apache License 2.0)
    - train time: around 12-18 hours

As our primary goal is on the retrieval-based data augmentation technique, we expect further optimization of task-specific model hyperparameters to improve performance. Note that our results are based on single runs, and running it multiple times with different random seeds may incur slight variation from the results we report.

B Privacy Policy Data Crawling & Retrieval Statistics

We crawl our English retrieval corpus from Google App Store using the Play Store Scraper[1]. In general, a privacy policy does not contain any personally identifiable information. However, there could be some mention of specific nomenclatures. There is no easy way to remove them, so we did not filter them manually. Note that we do not intend to use any commercial usage. However, below is the statistics of our (ERA) augmented corpus per each question category in the PrivacyQA training set.

Query Type	No. of Retrieval
Data Collection	2893
Data Sharing	1848
Data Security	891
Data Retention	542
User Access	145
User Choice	335
Others	14

Table 5: Retrieval statistics per query type.

C Effectiveness of Oracle Filtering and Different Top-k Selection

Method	Filter	top-k	Precision	Recall	F1
BERT-R	✗	10	39.9	50.8	44.7
	✓	10	46.5	45.5	46.0
PBERT-R	✗	10	48.4	45.6	46.9
	✓	10	46.9	43.3	45.1
	✗	50	47.8	45.5	46.7
	✓	50	49.5	46.3	47.8
SimCSE-R	✗	10	48.4	47.2	47.8
	✓	10	49.4	44.8	47.0
	✗	100	42.1	41.3	41.7
	✓	100	51.0	45.2	47.9

Table 6: Model performances with and without filtering with top-k. In general, without filtering, augmenting the retrieved candidates enhances recall but may reduce the precision (and hence may not improve the overall F1). Filtering, however improves the performance specially with larger top-k candidates.

[1] https://github.com/danieliu/play-scraper

D Qualitative Examples

Q: do you sell my photos to anyone?
Gold: i) We use third-party service providers to serve ads on our behalf across the Internet and sometimes on the Sites. (ii) These companies may use your personal information to enhance and personalize your shopping experience with us, to communicate with you about products and events that may be of interest to you and for other promotional purposes. iii) Your use of our Application with that healthcare institution may be subject to that healthcare institution's policies and terms. (iv) We may share personal information within our family of brands. (v) From time to time we share the personal information we collect with trusted companies who work with or on behalf of us. (vi) No personally identifiable information is collected in this process. (vii) We use third-party service providers to serve ads on our behalf across the Internet and sometimes on our Sites and Apps.
Correct Retrievals: (i) The Application does not collect or transmit any personally identifiable information about you, such as your name, address, phone number or email address. -(SimCSE-R) (ii) Some of this information is automatically gathered, and could be considered personally identifiable in certain circumstances, however it will generally always be anonymised prior to being viewed by Not Doppler, and never sold or shared. -(BERT-R) (iii) We also use the Google AdWords service to serve ads on our behalf across the Internet and sometimes on this Website. -(PBERT-R) (iv) To organ and tissue donation requests: By law, we can disclose health information about you to organ procurement organizations. -(BERT-R)
Incorrect Retrievals: (i) When you upload your photos to our platform or give us permission to access the photos stored on your device, your photo content may also include related image information such as the time and the place your photo was taken and similar "metadata" captured by your image capture device. -(SimCSE-R) (ii) These are not linked to any information that is personally identifiable.-(BERT-R)

Table 7: A fraction of retrieval examples (ii).

Q: who all has access to my medical information?
Gold: i) Apple HealthKit to health information and to share that information with your healthcare providers. ii) Your use of our Application with that healthcare institution may be subject to that healthcare institution's policies and terms.
Correct Retrievals: (i) We may share your information with other health care providers, laboratories, government agencies, insurance companies, organ procurement organizations, or medical examiners. -(SimCSE-R) (ii) Do not sell your personal or medical information to anyone. -(BERT-R) (iii) Lab, Inc will transmit personal health information to authorized medical providers. -(PBERT-R) (iv) To organ and tissue donation requests: By law, we can disclose health information about you to organ procurement organizations. -(BERT-R)
Incorrect Retrievals: (i) However, we take the protection of your private health information very seriously. -(SimCSE-R) (ii) All doctors, and many other healthcare professionals, are included in our database. -(PBERT-R) (iii) You may be able to access your pet's health records or other information via the Sites. -(BERT-R) (iv) will say "yes" unless a law requires us to disclose that health information.-(BERT-R) (v) do not claim that our products "cure" disease.-(BERT-R) (vi) Has no access to your database password or any data stored in your local database on your devices.-(BERT-R)

Table 8: A fraction of retrieval examples (i).