Retrieval Enhanced Data Augmentation for Question Answering on Privacy Policies

Md Rizwan Parvez[§], **Jianfeng Chi**[†], **Wasi Uddin Ahmad**[§],
Yuan Tian[§], **Kai-Wei Chang**[§]
[§]University of California, Los Angeles, [†]University of Virginia
[§]{rizwan,wasiahmad,yuant,kwchang}@ucla.edu, [†]jc6ub@virgina.edu

Abstract

Prior studies in privacy policies frame the question answering (QA) task as identifying the most relevant text segment or a list of sentences from a policy document given a user query. Existing labeled datasets are heavily imbalanced (only a few relevant segments), limiting the QA performance in this domain. In this paper, we develop a 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 cascade them with noise reduction filter models. Using our augmented data on the PrivacyQA benchmark, we elevate the existing baseline by a large margin (10% 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.

1 Introduction

Privacy policies describe how service providers collect, manage, and use their users' data. Understanding them is crucial for users as they can determine if the conditions outlined are acceptable. Policy documents, however, are lengthy, verbose, equivocal, and 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).

Building question answering (QA) systems for privacy policies is a stepping stone to allow users to ask questions about their rights. Prior works (Harkous et al., 2018; Ravichander et al., 2019) framed the QA task as a sentence selection task, essentially a binary classification task that identifies if a policy text segment is relevant to a question. Since policy documents consist of many sentences and typically a few are relevant to a question,[1] the classification data is imbalanced. In this work, we attempt to mitigate the data imbalance by augmenting positive QA examples. Specifically, we develop automatic retrieval models to supplement relevant policy sentences for each user query. We keep the queries unchanged as they are usually limited to a few forms only (Wilson et al., 2016).

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.

Unlike other domains, augmenting privacy policy statements is very challenging. First, they 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 (e.g., mixup (Zhang et al., 2018), back-translation (Edunov et al., 2018)). For

[1] PrivacyQA (Ravichander et al., 2019) dataset has 1,350 questions with an average number of answer sentences is 5, while the average length of policy documents is 138 sentences.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 201–210
May 2-6, 2023 ©2023 Association for Computational Linguistics

example, Kumar et al. (2020) identifies that even linguistically coherent instances augmented via generative models do not preserve the class labels well.[2] Hence, to reflect the data properties, we consider a retrieval-based approach to augment the raw policy statements. 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,[3] we then train a QA model (as a filter model) using the same pre-trained LM and data and couple it with the retriever.

Second, privacy policies are ambiguous; even skilled annotators dispute their diverse 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 sorts of relevant policy segments written in various diversified ways. To combat this insufficient data diversity, 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 filter models 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 (10% F1) and achieve a new one (50% F1). Furthermore, our ablation studies provide an insightful understanding of our model. We will release all data and code 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 outputs a binary label $z \in \{0, 1\}$ that indicates if q and p are relevant or not. As most sentences p are labeled 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 examples using filter models and aggregate final candidates. The final candidates are combined with the base data D to train the end QA model. We use an ensemble of retrievers and filter models built upon various pre-trained LMs throughout the whole process. Details are discussed in the following.

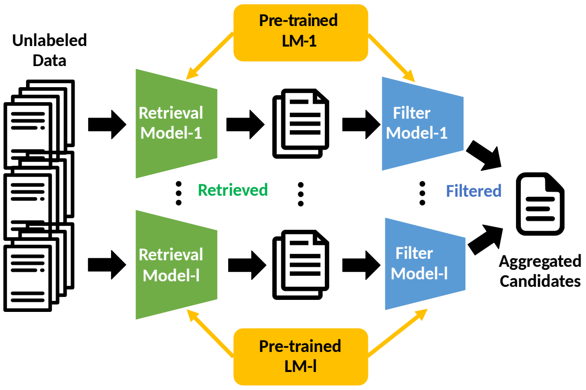


Figure 1: Our framework. Given a pre-trained LM, we train (i) a retriever, (ii) a QA model (filter) both on the small-size labeled data. From an unlabeled corpus, we first, retrieve the coarse relevant sentences (positive examples) for the queries in the training set and use the filter model to filter 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.

Retriever. Our retriever module is built upon the Dense Passage Retriever (DPR) model (Karpukhin et al., 2020). 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., $\text{sim}(q, p) = Q(q)^T \cdot P(p)$. 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 Model. To filter out the misclassified retrievals from $\mathcal{R}_L(\{q_i\}_{i=1}^m, \mathcal{P}, k)$, we train a QA (i.e., a text-classification) model ($\mathcal{Q}_L$) as a filter 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, the retriever model is a bi-encoder model that can pre-encode, index,

[2] To verify, in our preliminary study, for each positive example in the PrivacyQA training set, we augment a new synthesized positive example by en-zh-en back-translation using Google Translator API, and the performance drops by 3% F1.

[3] We refer the misclassified candidates as noisy retrievals.

and rank a large number of candidates. In contrast, our filtering model is a single cross-encoder model that can achieve comparatively higher performances (Humeau et al., 2019) (i.e., hence better as a filter) but can not pre-encode and hence can not be used for large-scale retrieval (more differences are below and in Appendix F). Once the retrieval is done, $\mathcal{Q}_L$ serves as an inexpensive binary classifier which naturally suits as a filter and fit into the pipeline. We verify the effectiveness of our filtering in Section 3.4. We denote retrieval outputs after filtering as $\mathcal{D}_L = \{(q,p,1) : \mathcal{Q}_L(q,p) = 1, \forall (q,p,1) \in \mathcal{R}_L(\{q_i\}_{i=1}^{m}, \mathcal{P}, k)\}$.

Training of $\mathcal{Q}_L$ and $\mathcal{R}_L$. Both the single encoder in $\mathcal{Q}_L$ and the two encoders in $\mathcal{R}_L$ are initialized with the same pre-trained LM L (e.g., BERT). Both are then fine-tuned as a binary classifier using the paired (query and policy is relevant or not) training data D. Additionally, to better-train the relatively weak bi-encoder $\mathcal{R}_L$, we consider the in-batch negative examples schema (Henderson et al., 2017; Parvez et al., 2021) and its hyper-parameters are tuned using mean ranking or mean reciprocal ranking (MRR) loss. At inference, for query q and candidate p, raw scores from $\mathcal{R}_L$ is used to rank and prediction {0,1} from $\mathcal{Q}_L$ is used to filter p.

Ensemble. In order to enhance the diversity and the quality of the retrieved candidates, we use a set of pre-trained LMs $\mathcal{L} = \{L_1, \dots, L_l\}$ 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

3.1 Setup

Evaluation Metrics. We evaluate our approach on PrivacyQA that is framed as a text classification task (Ravichander et al., 2019). We use *precision*, *recall*, and *F1 score* as the evaluation metrics.

Implementations. 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 6.5k 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.

Method	F	Precision	Recall	F1
Human	-	68.8	69.0	68.9
W/o data augmentation				
BERT+Unans.	-	44.3	36.9	39.8
BERT (reprod)		$48.0_{\pm 2.0}$	$37.7_{\pm 1.2}$	$42.2_{\pm 1.5}$
PBERT		$51.2_{\pm 0.4}$	$42.7_{\pm 0.6}$	$46.6_{\pm 0.4}$
SimCSE		$48.4_{\pm 0.8}$	$41.4_{\pm 0.7}$	$44.7_{\pm 0.7}$
Retriever augmented				
BERT-R	✗	$39.0_{\pm 0.8}$	$52.4_{\pm 1.7}$	$44.7_{\pm 0.4}$
	✓	$48.1_{\pm 1.4}$	$44.7_{\pm 0.9}$	$46.3_{\pm 0.5}$
PBERT-R	✗	$48.7_{\pm 1.9}$	$44.1_{\pm 1.8}$	$46.3_{\pm 1.6}$
	✓	$49.2_{\pm 1.6}$	$44.9_{\pm 2.0}$	$47.0_{\pm 1.2}$
SimCSE-R	✗	$47.0_{\pm 2.1}$	$44.5_{\pm 2.4}$	$45.7_{\pm 1.9}$
	✓	$48.6_{\pm 2.2}$	$43.9_{\pm 1.2}$	$46.1_{\pm 1.6}$
Ensemble retriever augmented				
Baseline-E	✗	$22.2_{\pm 0.8}$	$54.4_{\pm 0.8}$	$31.4_{\pm 0.8}$
ERA	✓	$47.4_{\pm 0.6}$	$50.5_{\pm 2.2}$	$\mathbf{48.9}_{\pm 0.8}$
ERA-D	✓	$51.0_{\pm 0.4}$	$48.7_{\pm 0.9}$	$\mathbf{49.8}_{\pm 0.7}$

Table 2: Test performances on PrivacyQA (mean$_{\pm\text{std}}$). F indicates filtering and BERT+Unans. refers to the previous SOTA performance (Ravichander et al., 2019). Retrieved candidates improve all the baseline QA models, especially when being filtered. Our ensemble retriever approach combines them and achieves the highest gains.

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. A previous implementation achieves the existing state-of-the-art performance (BERT+Unams. in Ravichander et al. (2019)). (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). Note that the retrieval database $\mathcal{P}$ is a subset of this data that is less noisy and crawled as a recent snapshot (more in Appendix E) (iii) *SimCSE*: We take the *PBERT* 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 filter model) and fine-tune on it the corresponding pre-trained LM as the final QA model. Finally, we

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	**49**

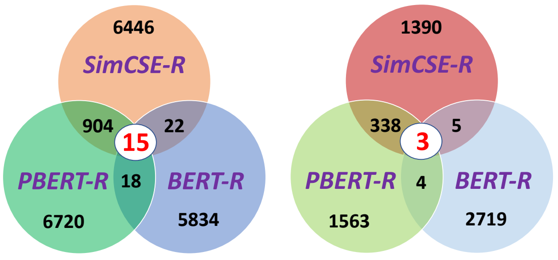
Table 3: F1-score breakdown (values are in Appendix D). B, PB, S refers to retrievers *BERT-R*, *PBERT-R*, and *SimCSE-R*. Different models performs better for different types (black-bold). ERA combines them and enhances performances for all categories (except: red).

consider one more ensemble retrieval augmented baseline (vii) *Baseline-E*, which is precisely the same as ours (settings below), except there are no intermediate filtering models.

Ours. We construct the augmented corpus $\mathcal{T}$ using (i) all 3 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 Augmentation* (ERA) and *Ensemble Retriever Augmentation–Domain Adapted* (ERA-D). By default, we fine-tune *SimCSE* as the final QA model.

3.2 Main Results

The results are reported 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, especially the recall score, as they are added as additional positive examples. However, these models may contain noisy examples (see Table 4), which lowers precision. Filtering these examples leads to improved precision for all retrievers. Finally, ERA and ERA-D aggregate these high-quality filtered policies–leading toward the highest gain (10% F1 from the previous baseline) and a new state-of-the-art result with an F1 score of ~50. Note that *Baseline-E* unifies all the candidates w/o any filtering performs considerably worse than all other models, including each individual retrieval model: *Baseline-E* augments more candidates as positives, which explains the highest recall score; in the meantime, as it does not filter any, the corresponding precision score is oppositely the lowest.



(a) (b)

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

3.3 Analysis

Table 3 shows the performance breakdown for different query types (more in Appendix D). For questions related to data collection, data sharing, and data security, the performance difference among the models is relatively small (≤ 5% F1); for data retention and user access, *BERT-R*, that is pre-trained on generic NLP texts, performs significantly well (> 15% F1), possibly because the answers to these query types focus on providing numerical evidence for the questions (e.g., How many days the data are retained?) that is less relevant to the domain of privacy policies; and for other types of questions the domain adapted models performs better (> 15% F1). Overall, the individual retrieval augmented models based on LM pre-trained w/ different corpora and objectives 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 policy statements describe similar information (i.e., have common phrases), they are often verbose and equivocal (i.e., multiple-different interpretations). Consequently, retrievers w/ different objectives and training corpora rank them differently. Therefore, although being retrieved from the same corpus, candidates retrieved by different models rarely match fully but may have notable overlapping information (words/phrases) and improve their performances equitably. For example, while the performances of *BERT-R*, *PBERT-R* and *SimCSE-R* w/ filtering are similar (~46) in Table 2, from Figure 2 their overlapping (exact match) is < 1% (qualitative examples in Appendix G). At the same time, their raw retrieval corpora have a high BLEU score of (≥ 0.78). This validates our hypothesis that retrievers built upon different pre-trained LMs learn distinct representations and hence retrieve diverse candidates.

3.4 Ablation Study

Sampling to tackle data imbalance. The preliminary experiments studied rebalancing techniques like equal sampling, but oversampling does not add new information and undersampling limits data, leading to poor generalization on unseen test data. Using equal positive and negative sub-sampled training instances resulted in a 9% drop in F1 score. Even augmented with a higher number of filtered positive examples retrieved by a single retriever model does not perform as well as when a lower number of ensemble-based positive instances are augmented. From Table 8 in Appendix, augmenting with a high number of filtered positive examples from a single retriever model performed worse than using a lower number of ensemble-based positive examples–suggesting the need for diverse and high-quality knowledge not present in training data.

A common filter. Performances of ERA (last row in Table 3) with a common filter model 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–showing that ERA is generic to end model choices.

Which pre-trained LMs to use? Table 3 shows ERA-D that combines fewer pre-trained LMs can 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 for future research (e.g., Parvez and Chang (2021)).

Recall performances on PrivacyQA dataset. Example retriever *BERT-R* scores the recall@k (k up to 10) values as 17, 28, 36, 42, 48, 53, 58, 59, 63, 67 respectively while the (cross-encoder) BERT QA model (i.e., filter model) achieves a recall (same as recall@1) of ~ 37. This shows the effectiveness of our designed filter model.

Can the final QA model be used as a filter and impact of filter models on the end performance? The final QA model can be used as a filter model. As for the single retriever model-based augmentation, the performance of the downstream end task depends on the performance of the retriever model, the filtering model, and the end QA model. A stronger filter model leads to better end QA performance in general. With a *BERT-R* retriever and *BERT* end QA model, the use of a *PrivacyBERT* filter model improves that of a *BERT* filter model from 46.3 to 46.8. However, when doing the ensembling using a combination of retriever and filter models from the same pre-trained language model (PLM) results in even better performance than using a stronger filter model from different PLMs.

Q: do you sell my photos to anyone?
Gold: i) We use third-parties to serve ads on our behalf across the Internet. (ii) We may share personal information within our family of brands. (iii) From time to time we share the personal information we collect with trusted companies who work with or on our behalf. (iv) No personally identifiable information is collected in this process.
Correct Retrievals: (i) *SimCSE-R*: The Application does not collect or transmit personally identifiable information such as your name, address, phone number or email address. (ii) *PBERT-R*: We also use the Google AdWords to serve ads on our behalf across the Internet and sometimes on this Website. (iii) *BERT-R*: To organ and tissue donation requests: By law, we can disclose your health information to organ procurement organizations.
Incorrect Retrievals: (i) *BERT-R*: These are not linked to any information that is personally identifiable. (ii) *SimCSE-R*: When you upload photos to our platform or give us permission to access the photos on your device, your photo content may also include related information such as the time and place your photo was taken and similar “meta-data” captured by your image capture device.

Table 4: Example retrieved policies. Retrieved candidates are distinct from expert annotated ones and can bring auxiliary knowledge to the model. Filtering is needed as inappropriate candidates can also be retrieved.

Qualitative examples. Table 4 shows some example retrievals of different models. Retrieved candidates are distinct from expert annotated ones and can bring auxiliary knowledge to the model.

4 Conclusion

We develop a noise-reduced retrieval-based data augmentation method that combines different pre-trained language models to address the data imbalance issue in privacy policy QA. However, our approach can possibly be adapted to other domains and we leave the exploration as the future work.

Limitations

In this work, we develop a retrieval-augmented QA framework specifically for Privacy Policies. Its effectiveness rooted on the characteristics of disjoint retrievals from different pre-trained language models (PLMs). Although this focused work completely aligns, addresses and adheres to the guidelines for a short-paper in this venue, we have not performed any experiments on data outside this privacy policy domain. Hence, the applicability of our method for any generic data domain is unknown. While the time/latency and resource utilization remains unchanged at inference-time, using multiple retriever modules our method introduce an additional overhead in model training. Using different implementation, PLMs, and random seeds may also lead to results that could be different from ours.

Ethics Statement

In this work, our approach crawls an unlabeled privacy policy corpus from the web policy documents specifically from the Google play store which we use as a retrieval database. Although these documents are completely publicly available and was used only for research purpose, they may contain some nomenclature of certain persons, objects, products, users, developers, or production houses (i.e., industries). We neither obfuscate nor make any altercation/modification to them.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work was supported in part by National Science Foundation Grant OAC 2002985, OAC 1920462 and CNS 1943100, Google Research Award, and Meta Research Award. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect those of the US Government or NSF.

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.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. 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*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

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.

Moniba Keymanesh, Micha Elsner, and Srinivasan Parthasarathy. 2021. Privacy policy question answering assistant: A query-guided extractive summarization approach. *arXiv preprint arXiv:2109.14638*.

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.

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.

Hoang Van, Vikas Yadav, and Mihai Surdeanu. 2021. Cheap and good? simple and effective data augmentation for low resource machine reading. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2116–2120.

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.

Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2020. Neural retrieval for question answering with cross-attention supervised data augmentation. *arXiv preprint arXiv:2009.13815*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

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 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). In relevant literature works, retrieval augmented methods are applied in various contexts including privacy policies (e.g., Van et al. (2021); Keymanesh et al. (2021); Yang et al. (2020)). Non-retrieval data aggregation has also been studied under different NLP contexts (e.g., bagging (Breiman, 1996), meta learning (Parvez et al., 2019)). However, we uniquely aggregate the retriever outputs using different pre-trained language models.

B 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 a particular set of models. For example, as mentioned in Section, 3.4, 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 8, 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 upto 4 runs using different random seeds.

C Privacy Policy Data Crawling & Retrieval Statistics

We crawl our English retrieval corpus from Google App Store using the Play Store Scraper[4].

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.

However, below is the statistics of our (ERA) augmented corpus per each question category in the PrivacyQA training set.

D PrivacyQA Dataset and Breakdown of Performance in Absolute Numbers

Table 6 shows the accuracy breakdowns in absolute numbers. In addition, A brief summary of the

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

Query Type	total	B	PB	S	ERA
Data Collection	6280	1901	1157	1186	1806
Data Sharing	4734	1332	777	1092	1268
Data Security	994	416	399	423	393
Data Retention	453	150	98	110	173
User Access	221	89	47	43	87
User Choice	493	91	49	24	55
Others	28	2	1	2	4
Overall	10332	3135	2084	2334	2935

Table 6: Number of correct predictions. Note that F1-score is not proportional to the accuracy. B, PB, S refers to retrievers *BERT-R*, *PBERT-R*, and *SimCSE-R*.

	PrivacyQA
Source	Mobile application
Question annotator	Mechanical Turkers
Form of QA	Sentence selection
Answer type	A list of sentences
# Unique policy docs	train: 27, test: 8
# Unique questions	train: 1350, test: 400
# QA instances	train: 185k, test: 10k
Avg Q. Length	train: 8.42 test: 8.56
Avg Doc. Length	train: 3.1k, test: 3.6k
Avg Ans. Length	train: 124, test: 153

Table 7: Brief summary of PrivacyQA.

PrivacyQA benchmark is in Table 7. Note that the questions in Privacy can be categorized into a few OPP-115 classes. These categories are enlisted in Table 3 in the main paper and the details of each category can be found in Wilson et al. (2016).

E Difference Between Pre-training and Retrieval Corpus

130k documents were collected before 2018 and by that time, the GDPR[5] and CCPA[6] were not enforced by then. Thus, the 130k documents are out-of-date and some content might not be comprehensive as the retrieval corpus. Besides, the 130k documents provided by (Harkous et al., 2018) contains some noises since we observe that the documents are not all written in English. However, as the data size is larger, we still use it for pre-training. In contrast, our corpus was collected after 2020 and we filtered out some possible noises (e.g., filtering out non-English document) while crawling.

F Difference Between the Filtering Model and the Retriever

The retriever model is a bi-encoder model whose model parameters are fine-tuned with in-batch negative loss (discussed in Section 2 in the main paper), hyper-parameters are tuned based on average rankings (`https://github.com/facebookresearch/DPR/blob/a31212dc0a54dfa85d8bfa01e1669f149ac832b7/train_dense_encoder.py#L294`) and that can pre-encode, index and rank a large number of candidates while our filtering model is a cross-encoder text-classifier (e.g., single encoder fine-tuned BERT) that is fine-tuned w/o any additional in-batch negatives and in-general achieves comparatively higher performance (Humeau et al., 2019) (i.e, better as a filter) but can not pre-encode and hence can not be used for large scale retrieval.

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 8: Model performances with and without filtering (i.e., w/o the filter model) 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, especially with larger top-k candidates. In above, top-100 augmented examples (~13K total positives after filtering) retrieved by a single retriever perform worse than with top-10 examples by ERA or ERA-D (total 7K or 4K positive examples) reported in Table 2 in the main paper.

[5] https://gdpr-info.eu/

[6] https://oag.ca.gov/privacy/ccpa

G More Qualitative Examples

The below tables show some example retrievals of different models. Retrieved candidates are distinct from expert annotated ones and can bring auxiliary knowledge to the model.

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 9: A fraction of retrieval examples (i).

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 10: A fraction of retrieval examples (ii).