Privacy and Customer's Education: NLP for Information Resources Suggestions and Expert Finder Systems

Luca Mazzola[1][0000−0002−6747−1021], Andreas Waldis[1][0000−0002−2772−5701], Atreya Shankar[1], Diamantis Argyris[1], Alexander Denzler[1], and Michiel Van Roey[2]

[1] HSLU - Lucerne University of Applied Sciences and Arts;
School of Information Technology,
Suurstoffi 1, CH-6343, Rotkreuz, Switzerland
{luca.mazzola, andreas.waldis, atreya.shankar, diamantis.argyris, alexander.denzler}@hslu.ch
[2] Profila GmbH,
Seeburgstrasse 45, CH-6006, Luzern, Switzerland
info@profila.com

Abstract. Privacy is one of the key issues for citizen's everyday online activities, with the United Nations defining it as "a human right in the digital age". Despite the introduction of data privacy regulations almost everywhere around the globe, the biggest barrier to effectiveness is the customer's capacity to map the privacy statement received with the regulation in force and understand their terms. This study advocates the creation of a convenient and cost-efficient question-answering service for answering customers' queries on data privacy. It proposes a dual step approach, allowing consumers to ask support to a conversational agent boosted by a smart knowledge base, attempting to answer the question using the most appropriate legal document. Being the self-help approach insufficient, our system enacts a second step suggesting a ranked list of legal experts for focused advice. To achieve our objective, we need large enough and specialised dataset and we plan to apply state-of-the-art Natural Language Processing (NLP) techniques in the field of open domain question answering. This paper describes the initial steps and some early results we achieved in this direction and the next steps we propose to develop a one-stop solution for consumers privacy needs.

Keywords: Privacy · Data Privacy Regulation · Natural Language Processing · Consumers' privacy · Read and Retrieve in Open Domain Question Answering

1 Introduction

Privacy is one of the key issues for citizen's everyday online activities. Online Privacy Literacy (OPL) is a recent approach to estimate knowledge about privacy

rights, considering declarative an procedural aspects of preventive and corrective protective strategies [14]. The most important finding in this study, is the fact that an increase in theoretical users' knowledge does not reflect a reduction of their concerns, but rather ends up in an increased interest in understanding and fully exploiting the protective measures offered by relevant laws and regulations. This is also reflected in the higher consideration given to privacy protection from government and non-governmental institution, such as the United Nations [10], defining it as "a human right in the digital age" in the context of the current pervasive datafication. Despite the introduction of data privacy regulations almost everywhere around the globe, the biggest barrier to effectiveness is the customer's capacity to map the privacy statement received with the regulation in force and understand their terms.

The matter is twofold: on one side this is a complete switch of paradigm from the social network approach, where a service is offered to people free-of-charge because the business model came from the usage of consumers' personal information. n the other side, the content of companies' privacy policies and notices (at least, at the European level) are mainly guided by the GDPR; which is a legislative text and therefore uses typical juridical organisation of information and legal jargon, which makes comprehension by non-experts difficult. These two factors produce a barrier when a consumer would like to decide if a specific privacy-related contract is compatible with their desired level of privacy. Despite some analyses of GDPR effects such as the *"right to explanation"* [23], no major developments have concentrated on the education and explanation of privacy terms for general consumers, but only on the raised awareness. This mismatch between the risk realisation and the practical applicability of the tools provided by the legal framework has brought a sense of frustration for consumers as described in [20] with comments such as *"[the law] is not clear and simple to me while you come across it all the time and it impacts your data"*. As such, there is a need for immediate consumers' support while dealing with personal-data collection, storage and usage, in order to provide remediation actions for this tension and, thus, a more relaxed interaction with personalisation approaches. Figure 1 gives a graphical overview of this initial step.

2 Idea

While interacting with websites or registering for services, customers usually encounter privacy contracts stating the terms of agreement. Due to the complexity of these documents, customers usually accept these terms without understanding them completely. Our objective is to support customers' comprehension of such privacy contracts. By helping the average customer in understanding these contracts, we hope to foster more awareness regarding data privacy and individual rights regarding personal data usage. A natural interaction with the knowledge base is a precondition for the a positive usability by the average citizen. Conversational agents, that enhances traditional chatbots by adding contexts and users' goals, are well-known for their ability to guide the user towards the required information.

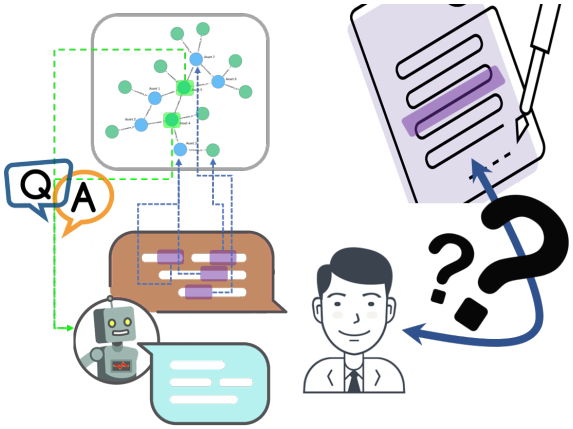

Fig. 1. Project graphical abstract: a customer is dealing with a privacy related agreement online and cannot fully understand the terms and conditions. Before approving it, they would like to clarify some aspects, using a conversational agent, that can point to the better suited document in a Q&A collection.

They were already adopted successfully in context such as cybersecurity [9], health [13], and agriculture [11].

As part of this objective, our idea is to develop a smart Knowledge Base (KB) composed of documents that can be matched with the specific question a user has. This will work as the intelligence behind a simple conversational agent which we propose as a self-help tool. This should be able to solve the most common and standard questions, but will likely fail for more specific needs.

Figure 2 presents the case where the smart KB alone is inadequate in resolving customer doubts **(a)**, either by not finding a fit-enough answer in the Q&A or by receiving a negative feedback from the customer on the provided answers set. In this situation, the system can propose to provide suggestion for legal scholars that can analyse the remaining open points and provide the user with a personalised answer. Specifically, the same knowledge base **(b)** will be employed to find matches between the question and human experts' profiles **(c)** resulting in a ranked list of specialised legal scholars and their rates **(d)**, amongst which the customer can choose to receive a tailor-made interpretation **(e)**. This can provide a rapid and low-cost option to clarify specific questions. Additionally, user feedback regarding the perceived quality of answers and experts suggested will be collected to internally improve performance and experts' profiles. Our final vision is to provide a one-stop-solution for every user's privacy-oriented questions.

The creation of the smart KB and its matching algorithm will be based on Natural Language Processing (NLP) in order to support the adoption of natural language in the interaction with the user, and to naturally support a conversational agent interface.

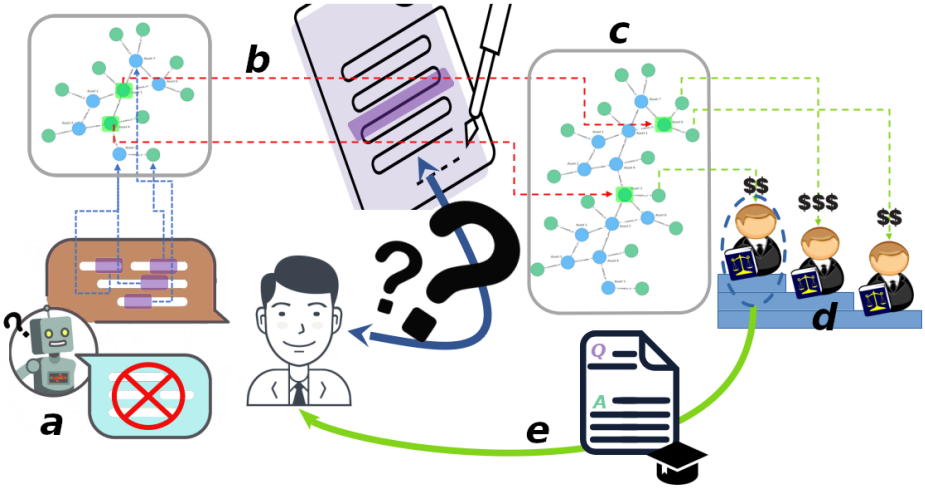

Fig. 2. Expert support: a request to the conversational agent is failed or not satisfactory **(a)**, so the system uses the smart KB **(b)** to checks amongst the available experts **(c)** a pool of suitable professionals to answer the specific question, considering both its depth and breath. As results, a ranked list of professional is returned to the consumer together with the fees required by each of them **(d)**. The customer can then select the preferred one, pay the fees and receive the qualified answer to his question **(e)**. The answer is then stored and its evaluation is used to update the expert's profile for the following interactions.

Requirements Elicitation

The main aspects such a one-stop solution should provide were identified as from the following list, that will be refined and extended during the project development:

1. The knowledge base (KB) should include documents of different origin and scope, and should consider their legal relevance and normative strength in the analysis. The main identified categories are:
   (a) Legislative acts (at different level: international, European, national)
   (b) Juridical acts (Court & Administrative cases)
   (c) Regulatory (such as guidelines and recommendations advises from Data Protection Agencies (DPA))
   (d) Consumer organisations
   (e) Online Publications of Legal Research (both professional and non-professional)
   (f) Company Policies
2. An abstraction layer should be included in the KB, to recognise concepts included in documents and natural language statements and their relationships, thus homogenising their internal representation
3. The abstraction layer should provide a granular approach [19], in order to allow the joint consideration of specificity (depth) and coverage (breath) of a given input

4. Should be possible to extend the smart KB, by adding/updating data at all levels.
5. The underlying technology for the smart KB should offer a standardised and repeatable automated way to regenerate it, and should require minimal human intervention in the process (thus banning any rule-based approach for document analysis)
6. The produced inference process on users input should privilege the run-time efficiency, to allow scalability of the solution. The offline KB update could be more computationally expensive, but should be possible to run it without putting offline the service
7. The newly produced legally qualified content by human experts should be indexed and be available for the conversational agent, thus extending dynamically and focused the coverage of the KB
8. The KB should be usable to profile legal scholars based on their specialisation and to rank them as the better suited to answer a specific user request, based on the matching on depth and breath of the knowledge required fro answering with the expertise offered by the human expert

Additionally, the KB will be initially developed using the English language only, but should be possible to include multilingualism in a further stage. This is important for integration of European privacy-regulations, but also for countries with more than a single official language, such as Switzerland, were we plan to start from.

Based on these requirements, the following steps were taken, as described next. Sect. 3 report the results of the manual data collection, whether in Sect. 4 the automatic process of company privacy policy gathering and validation is described. Those two parts form the base for the initial KB and abstraction layer creation, as from the requirements [1-3]. Regarding requirements [5-6], Sections 5 and 6 provide some initial exploration and potential directions for further research. Sect. 7 depicts some of the open issue still to be solved to provide a one-stop solution for a privacy-related expert finder. Eventually, the steps already taken and the future direction towards a usable system are sketched in Sect. 8, that concludes our contribution.

3 Manual Data Collection

AS stated before, currently we focus only on English based documents. The first part of the data collection was performed manually by experts. This part concentrates equally on all categories listed within the Requirement 1, with a particular attention in having an overall coverage of all levels. Table 1 resume the result of the manual data collection process.

[1] These policies appear in the website of the companies, and they represent a manually selected set of highly relevant companies operating on the Swiss/European market, in particular Germany, Austria and Switzerland (DACH).

ID	Category	#Subcategory	#Documents (EN)
(a)	Legislation	9	78
(b)	Juridical acts	9	189
(c)	Regulatory	6	123
(d)	Consumer organisations - guidelines	5	48
(e)	Online Legal Research	6	209
	$\gg$ (e.1) *Professional*	*3*	*132*
	$\gg$ (e.2) *Non-professional*	*3*	*77*
(f)	Privacy Company Policies[1]	4	256
	TOTAL	39	**903**

Table 1. Manual data collection results, as from requirement [1] in Sect. 2.

An observation is that category (a) has a finite number of documents, being legislative acts privacy-relevant a bounded set, given a specific scope and language. Juridical acts (b) increases in time, but in a slow way, being official acts that requires a significant official tribunal involvement. Thus, at a specific moment in time, the coverage that can be guaranteed in the manual data collection appears to be sufficient. A similar reasoning applies for Regulatory (C) and Consumer Organisation Guidelines (d). Category (e) of Online Legal Research (professional and non), these are User Generated Content (UGC) and we plan to not rely too much on them, at least at the beginning. Eventually, for Privacy Company Policies (f) an initial set of highly relevant document was compiled, retrieved and saved by hand in the project repository, but due to the cost of this data collection, an automatic process was implemented, as described in the next section.

4 Company Privacy Policies

As part of our approach in building a smart KB, we ascertained a need to develop a larger database of company privacy policies. This database could be used for several purposes, such as training language models and learning privacy-related concepts in real-world data, that will not be possible on the amount of privacy policy collected with the manual approach ($\sim$250) . Upon literature review, we found a few studies which developed a similar privacy policy database [1,2,15]. [2] is a relevant study which used historical data from the *Wayback Machine* to crawl privacy policy data. While useful, our objective required a slightly different approach since we require recently crawled privacy policies rather than historical ones.

Adapting our approach from [2], we first needed to identify a catalogue of companies with their respective web domains, headquarter locations and rough

sizes. While [2] used the Alexa Rank[2] to identify domains for crawling, we decided to use the 7+ Million Company data set[3] to start our crawling process. This was mainly because the 7+ Million Company data set offered metadata on companies that other sources did not. Upon retrieving this data set, we filtered the entries and kept company data where web domains and headquarter locations were available. Following this initial filtering, we crawled company domains using the following two distinct approaches.

4.1 Raw Crawl

We coin the crawling first approach as *Raw Crawl*. For this approach, we visited each company's domain and searched the raw web pages for certain regular expressions. In case this page was in the English language, we searched for terms such as *"privacy"* and *"data protection"*. In case this page was not in English, we used separate regular expressions, such as *"en"* and *"eng"*, to attempt to find an English language version of the page. All scraped results were saved in a SQLite database.

4.2 Search Engine

We coin the second crawling approach as *Search Engine*. Here, we utilised the crawling power of search engines in order to find respective privacy policies. Specifically, we used the DuckDuckGo[4] search engine API to search for and retrieve company's privacy policies. This had an added benefit of providing cleaner data. Similar to the first approach, we saved all scraped results in a SQLite database.

4.3 Policy Classifier

While crawling and scraping data from the web, it is common to encounter noise in the form of unintended documents and languages. While there exist

[2] https://www.alexa.com/topsites
[3] https://www.kaggle.com/peopledatalabssf/free-7-million-company-dataset
[4] https://duckduckgo.com/

Hyperparameter	Value	Description
Minimum DF	0.1	Minimum document term frequency for consideration
N-gram range	1-4	Word-level N-gram ranges used as inputs
Maximum depth	15	Maximum depth of each decision tree
Minimum leaf samples	3	Minimum number of samples at each leaf node
Estimators	200	Number of decision trees in the forest

Table 2. Hyperparameters used for our Random Forests privacy policy classifier

Approach	# Companies	# Policies	# Policies [en]	# Filtered policies [en]
Raw Crawl	655'374	746'345	634'950	475'726
Search Engine	5'404	5'404	5'355	3'057
TOTAL	660'778	751'749	640'305	478'783

Table 3. Summary statistics from our two crawling strategies; "en" refers to English language documents; filtered policies refer to policies which exceeded a classifier probability threshold of 0.75

several ways to mitigate such noise, we decided to follow a pre-tested approach from [2]. This approach involves utilising a Random Forests (RF) privacy policy classifier, which essentially classifies a given document as a privacy policy or not given a certain confidence threshold. Similar to [2], we trained a Random Forests ensemble classifier [3] on annotated privacy policies retrieved from the study's publicly available data sources. For training this classifier, we utilised word-level TF-IDF features and hyperparameters as mentioned in Table 2. Our best privacy policy classifier achieved a precision and recall of 98.9% and 89.5% respectively at a probability threshold of 0.75. We used this aforementioned classifier threshold to partition our scraped data such that we only kept privacy policies of high quality.

4.4 Manual evaluation

In order to test the RF classifier, and our assumption that only good quality privacy policy will get accepted with this setting, we performed a short manual evaluation. First we created for classes of outcome, based on the achieved probability closeness to the threshold:

- STRONG REJECT: documents with probabilities significantly inferior to the threshold ($p \ll 0.75$)
- WEAK REJECT: documents not in the previous class and presenting probabilities inferior but close to the minimum level ($0.75 - \epsilon < p < 0.75$)
- STRONG ACCEPT: documents with probabilities significantly superior to the threshold (close to the unit) ($p \gg 0.75 \implies p \simeq 1$)
- WEAK ACCEPT: documents not in the previous class and presenting probabilities superior but close to the minimum level ($0.75 \leq < p < 0.75 + \epsilon$)

We then sample 14 documents for each category and asked four users (u1 .. u4) to evaluate the documents as valid privacy policy (1) or not (0). The results of this evaluations are reported in the Table 4. This result clearly shows a strong agreement between the annotators themselves, as well as between the annotators and the RF model.

Category	evaluation	U1	U2	U3	U4
STRONG REJECT	1	0	0	0	0
	0	14	14	14	14
WEAK REJECT	1	0	1	1	3
	0	14	13	13	11
WEAK ACCEPT	1	14	14	14	14
	0	0	0	0	0
STRONG ACCEPT	1	14	14	14	14
	0	0	0	0	0

Table 4. Classifier manual evaluation: all the documents accepted by the RF are also accepted by the human evaluators, while some of the documents (close to the probability threshold) will be rejected even if a human user could accept them as valid.

4.5 Results

Table 3 shows the results of our two crawling approaches. We utilised the *Raw Crawl* approach more frequently than the *Search Engine* approach due to rate limiting from the latter which slowed down our overall crawls. In total we were able scrape ∼660K companies and obtain ∼480K filtered English-language privacy policies; specifically those whose classifier probabilities exceeded our preset threshold of 0.75. In addition, we were able to obtain privacy policies from various European languages such as Dutch, French and German as a side-effect of our crawling approaches. These additional languages could potentially be used for downstream multi-lingual privacy analysis tasks.

5 Question Answering

Another essential aspect of our smart KB involves gathering real-world questions and answers related to privacy. This data could be used to train and evaluate our language models in the legal privacy domain. To do this, we first identified two important data sources for questions and answers; namely Law Stack Exchange and Reddit. In the next subsections, we describe these sources and our crawling approaches further.

5.1 Law Stack Exchange

Law Stack Exchange[5] is a subset of Stack Exchange where questions and answers are limited to the legal domain. We found this to be a good source of questions and answers for our smart KB. To crawl this data source, we simply utilised the

[5] https://law.stackexchange.com/

Source	# Questions	# Answers	Answers per question
Law Stack Exchange	1'349	1'988	1.47
Reddit	45'742	190'666	4.17
TOTAL	47'091	192'654	4.09

Table 5. Summary statistics from our Law Stack Exchange and Reddit question answering data

Stack Exchange Data Dumps[6] and extracted data containing any of the *"gdpr", "privacy", "data-protection", "data-ownership", "ccpa", "confidentiality", "coppa", "ferpa", "can-spam-act-of-2003"* and *"tcpa"* tags into a SQLite database.

5.2 Reddit

Reddit is another source of useful questions and answers in the privacy domain. Based on our analysis, we found the *"gdpr", "privacy", "europrivacy", "privacylaw", "netneutrality"* and *"eff"* subreddits to be most useful for our smart KB. We utilised `pushshift.io`[7] to retrieve submissions and comments from the aforementioned subreddits and extracted this information into the same SQLite database.

5.3 Results

Table 5 shows a summary of the questions and answers extracted from Law Stack Exchange and Reddit. We can observe that Law Stack Exchange has fewer overall questions and answers, as well as answers per question, compared to Reddit. The difference could be attributed to Reddit being a more conversational platform with several exchanges per question compared to Law Stack Exchange being more formalised.

5.4 Unsupervised ML Evaluation

Before fine-tuning language models on our question answering data, we ran unsupervised evaluation on these questions and answers using pre-trained language models. This would give us a baseline as to how well various language models perform. For simplicity, we performed unsupervised evaluation only on the Law Stack Exchange data set. Our methodology for unsupervised evaluation is as follows:

1. Encode all questions and answers based on the specific model to get vector representation with 768 dimensions

[6] https://archive.org/details/stackexchange
[7] https://github.com/pushshift/api

Language Model	Evaluation Mode	K-candidates	Accuracy
SBERT (`all-mpnet-base-v2`) [16]	Local	1	**0.844**
		2	0.979
		3	0.991
		4	0.999
		5	1.000
	Global	1	**0.607**
		2	0.746
		5	0.838
		10	0.896
		20	0.931
Legal-BERT (`legal-bert-base`) [6]	Local	1	0.815
		2	0.973
		3	0.991
		4	0.997
		5	0.999
	Global	1	0.086
		2	0.116
		5	0.166
		10	0.210
		20	0.262

Table 6. Unsupervised question answering evaluation for Law Stack Exchange

2. Calculate the cosine similarity of each question-answer pair
3. Evaluate local (find best answer from one thread) and global (find best answer from all answers of the data set) by selecting a set of K candidate answers
4. Calculate the accuracy depending on whether the target answer is within the first K candidates

Table 6 shows a summary of unsupervised evaluations for SBERT [16] and Legal-BERT [6]. Here, we observe that SBERT outperforms Legal-BERT for both the local and global evaluation modes. This also makes sense since SBERT is trained to produce reasonable sentence representation and is partly fine-tuned on QA data. This provides us with interesting insights, since we simultaneously require the legal language understanding of Legal-BERT and the question-answering capacity of SBERT. A combined approach towards our smart KB will likely involve combining the training approaches for SBERT and Legal-BERT.

Model	ECtHR-A	ECtHR-B	SCOTUS	EUR-LEX
Legal-BERT (Small) [6]	**0.626**	**0.694**	**0.597**	0.482
Distil-BERT (Base) [17]	0.611	0.691	0.559	**0.515**
Mini-LM [24]	0.551	0.610	0.455	0.356
BERT-Tiny [22]	0.440	0.504	0.357	0.250

Table 7. Macro-F_1 evaluation scores for ECtHR-A, ECtHR-B, SCOTUS and EUR-LEX

Model	LEDGAR	UNFAIR-ToS	CaseHOLD
Legal-BERT (Small) [6]	**0.820**	**0.817**	**0.729**
Distil-BERT (Base) [17]	0.815	0.794	0.686
Mini-LM [24]	0.796	0.132	0.713
BERT-Tiny [22]	0.733	0.111	0.662

Table 8. Macro-F_1 evaluation scores for LEDGAR, UNFAIR-ToS and CaseHOLD

5.5 Potential Bias

As per Table 6, we observed that SBERT had a significantly higher global unsupervised performance than Legal-BERT. To investigate causes, we looked into potential biases in the input data that could assist SBERT. We found that several answers had parts of the question quoted in them. We suspected these quotations to be forms of bias that could have contributed to the high performance of SBERT. An open task in question answering is to investigate how strongly the quotes in answers bias existing language models.

6 LexGLUE

An important part of developing our smart KB is to develop appropriate benchmarks for evaluation. Recent developments in NLP show a shift towards multi-task benchmarks for evaluating language models. As our focus is in the legal and privacy domain, we decided to use the LexGLUE benchmark from [8] as a starting point for benchmarking our smart KB. LexGLUE consists of 7 English language tasks from the legal domain; namely ECtHR-A [4], ECtHR-B [7], SCOTUS [18], EUR-LEX [5], LEDGAR [21], UNFAIR-ToS [12] and CaseHOLD [25].

Due to limited resources, we decided to start testing small and medium-sized Transformer language models on the LexGLUE benchmark. We envisioned these models as potential candidates for our smart KB. Tables 7 and 8 summarise the

results of four smaller-sized models on the LexGLUE benchmark. Here, we can observe the Legal-BERT (small) performs the best on all tasks except EUR-LEX. After computing these scores, we reported these test results to [8] in order to support their benchmarking of multiple models.

While LexGLUE represents a comprehensive benchmark for evaluating our smart KB, we strongly believe in the creation of more appropriate benchmarks for our use-case. Since our smart KB is envisioned to perform open domain read-and-retrieve along with basic reading comprehension tasks, we would need to augment our evaluation benchmark with such tasks.

7 Expert Finder

As described in Sect. 2, the smart KB can cover the most general cases, where an answer is already present and the level are satisfactory for the consumer's needs. Anyway, when the system is unable to provide an answer or the user is not satisfied with the focus or precision of the reported resources, a second level support will be proposed to the user, involving the expertise of legal scholars, and some personalised fees to be paid for the service. In order to provide the best match (not always the most expert person in the subject asked, but with the right combination of proficiency focus and depth) the proposed solution will use again the granular knowledge base to map contributions of participating layers into juridical profiles comparable with questions characterisation. Thus, the same smart KB can be adopted to overcome the major barriers present in current expert finder systems, namely:

1. Explicit declaration of skills and knowledge by experts can be biased
2. Lack of consideration for different areas of expertise and different levels of knowledge
3. Classification of questions with regard to content has to be done manually and is a very complex/time-consuming task for users
4. Knowledge evolves over time, and capturing this evolution explicitly is problematic
5. Difficulty in determining which legal domain a question fits in, leading to sub-optimal forwarding of questions to experts
6. The cost of answering a question is fixed instead of being related to the complexity of the question

Our system will solve those issue, by moving the experts profile creation from an explicit, manually-declared process to an implicit, automatically-tracked approach. On top of it, time evolution will be considered, as newly provide answers from legal scholars will be used to align their internal characterisation in terms of both expertise focus and depth. Eventually, consumer's feedback will be used to build a reputation system, that will cooperate with the expert profiles to determine their positioning in the ranked list suggested to the user.

8 Conclusion

This positional paper advocates the creation of a convenient and cost-efficient question-answering service for answering customers' queries on data privacy. It proposes a dual step approach: first by developing a conversational agent supported by a smart knowledge base which attempts to answer the question using the most appropriate legal document. In case the first step is insufficient, our system enacts a second step and suggests a ranked list of legal experts for focused advice. All of the matching will be supported by a granular knowledge base, enabling semantic matching between consumer's questions and privacy related documents or legal scholars providing the second-level support under the payment of fees personalised based on the specificity and coverage of the asked support.

To create the smart KB, after identifying some initial long-term requirements, we started classifying and collecting relevant documents. After manually retrieving most part of the stable and accessible sources, we turned to crawling and scraping companies' privacy policies along with real-world questions and answers from Law Stack Exchange and Reddit, to create a large enough dataset to be usable for Machine Learning approaches. We have also performed unsupervised ML evaluation on the Law Stack Exchange, which showed promising results for SBERT. Finally, we ran the LexGLUE benchmark on small and medium-sized language models. The results of this benchmark showed promising results for Legal-BERT. We determined that our smart KB would need to utilise training procedures from both SBERT and Legal-BERT.

To create our expert finder system, we first identified significant issues that we would need to overcome. With these open issues listed, we will start developing the expert finder in the next steps.

Acknowledgements

The research leading to this work was partially financed by *Innosuisse* - Swiss federal agency for Innovation, through a competitive call. The project 50446.1 IP-ICT is called *P2Sr Profila Privacy Simplified reloaded: Open-smart knowledge base on Swiss privacy policies and Swiss privacy legislation, simplifying consumers' access to legal knowledge and expertise.*[8]. The authors would like to thanks all the people involved on the implementation-side at Profila GmbH[9] for all the constructive and fruitful discussions and insights provided about privacy regulations and consumers' rights.

References

1. Ahmad, W., Chi, J., Tian, Y., Chang, K.W.: PolicyQA: A reading comprehension dataset for privacy policies. In: Findings of the Association for Computational

[8] `https://www.aramis.admin.ch/Grunddaten/?ProjectID=48867`
[9] `https://www.profila.com/`

Linguistics: EMNLP 2020. pp. 743–749. Association for Computational Linguistics, Online (Nov 2020), https://www.aclweb.org/anthology/2020.findings-emnlp.66

2. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., Mayer, J.: Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In: Proceedings of The Web Conference 2021. p. 22. WWW '21, Association for Computing Machinery (2021). https://doi.org/10.1145/3442381.3450048, https://doi.org/10.1145/3442381.3450048
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Chalkidis, I., Androutsopoulos, I., Aletras, N.: Neural legal judgment prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4317–4323. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1424, https://aclanthology.org/P19-1424
5. Chalkidis, I., Fergadiotis, M., Androutsopoulos, I.: Multieurlex - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. CoRR **abs/2109.00904** (2021), https://arxiv.org/abs/2109.00904
6. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2898–2904. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.261
7. Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., Malakasiotis, P.: Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 226–241. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.22, https://aclanthology.org/2021.naacl-main.22
8. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: Lexglue: A benchmark dataset for legal language understanding in english. CoRR (2021), arXiv: 2110.00976
9. Franco, M.F., Rodrigues, B., Scheid, E.J., Jacobs, A., Killer, C., Granville, L.Z., Stiller, B.: Secbot: A business-driven conversational agent for cybersecurity planning and management. In: 2020 16th International Conference on Network and Service Management (CNSM). pp. 1–7. IEEE (2020)
10. Gstrein, O.J., Beaulieu, A.: How to protect privacy in a datafied society? a presentation of multiple legal and conceptual approaches. Philosophy & Technology **35**(1), 1–38 (2022). https://doi.org/10.1007/s13347-022-00497-4
11. Jain, M., Kumar, P., Bhansali, I., Liao, Q.V., Truong, K., Patel, S.: Farmchat: a conversational agent to answer farmer queries. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(4), 1–22 (2018)
12. Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H., Sartor, G., Torroni, P.: CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. CoRR **abs/1805.01217** (2018), http://arxiv.org/abs/1805.01217
13. Meier, P., Beinke, J.H., Fitte, C., Behne, A., Teuteberg, F.: Feelfit - design and evaluation of a conversational agent to enhance health awareness. In: Krcmar, H., Fedorowicz, J., Boh, W.F., Leimeister, J.M., Wattal, S. (eds.) Proceedings of the 40th International Conference on Information Systems, ICIS 2019, Munich,

Germany, December 15-18, 2019. Association for Information Systems (2019), https://aisel.aisnet.org/icis2019/is_health/is_health/22
14. Prince, C., Omrani, N., Maalaoui, A., Dabic, M., Kraus, S.: Are we living in surveillance societies and is privacy an illusion? an empirical study on privacy literacy and privacy concerns. IEEE Transactions on Engineering Management pp. 1–18 (2021). https://doi.org/10.1109/TEM.2021.3092702
15. Ravichander, A., Black, A.W., Wilson, S., Norton, T., Sadeh, N.: 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). pp. 4949–4959. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1500, https://www.aclweb.org/anthology/D19-1500
16. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084
17. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019), http://arxiv.org/abs/1910.01108
18. Spaeth, H., Epstein, L., Ruger, T., Whittington, K., Segal, J., Martin, A.D., Ruger, T.J., Benesh, S.C.: Supreme court database code book. URL: https://scdb.wustl.edu (2020)
19. Stalder, F., Denzler, A., Mazzola, L.: Towards granular knowledge structures: Comparison of different approaches. In: 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 261–266. IEEE (2021)
20. Strycharz, J., Ausloos, J., Helberger, N.: Data protection or data frustration? individual perceptions and attitudes towards the gdpr. Eur. Data Prot. L. Rev. **6**, 407 (2020)
21. Tuggener, D., von Däniken, P., Peetz, T., Cieliebak, M.: LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1235–1241. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.155
22. Turc, I., Chang, M., Lee, K., Toutanova, K.: Well-read students learn better: The impact of student initialization on knowledge distillation. CoRR **abs/1908.08962** (2019), http://arxiv.org/abs/1908.08962
23. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
24. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. CoRR **abs/2002.10957** (2020), https://arxiv.org/abs/2002.10957
25. Zheng, L., Guha, N., Anderson, B.R., Henderson, P., Ho, D.E.: When does pretraining help? assessing self-supervised learning for law and the casehold dataset. CoRR **abs/2104.08671** (2021), https://arxiv.org/abs/2104.08671