A Question Answering Tool for Website Privacy Policy Comprehension

Luca Mazzola[1](✉), Atreya Shankar[1], Christof Bless[1], Maria A. Rodriguez[1], Andreas Waldis[1], Alexander Denzler[1], and Michiel Van Roey[2]

[1] School of Information Technology, HSLU - Lucerne University of Applied Sciences and Arts, Suurstoffi 1, 6343 Rotkreuz, Switzerland
`{luca.mazzola,atreya.shankar,christof.bless,maria.anduezarodriguez, andreas.waldis,alexander.denzler}@hslu.ch`

[2] Profila GmbH, Seeburgstrasse 45, 6006 Luzern, Switzerland
`info@profila.com`

Abstract. Everyday we interact with online services from companies that ask for our permission to use our personal information. Nowadays it is common practice for websites and apps to collect big amounts of data which are mainly used for revenue optimization based on user analytics. This customer data collection and usage is regulated by legal agreements (i.e., privacy and cookie policies) which we are required to accept (multiple times a day), but which are generally very long and formulated in a way that makes their interpretation difficult for the general public. An average privacy policy takes 15 min to read and includes lots of legal jargon (e.g., including words like "data controller" and "legal basis for processing"). In this research project, we are developing a support system where users can search for concrete answers in the privacy policies of companies or websites, by formulating their questions in natural language. Instead of blindly accepting a privacy policy, a user could first query the system for answers to a potential concern. The system will return a ranked list of phrases and documents matching the query. In case the generated answer is not sufficient for the user, an extension will allow them to forward complex requests to best-matching legal professionals, specialized in privacy legislation, which can process them for a small fee. We present different aspects of the internal implementation, including the identification of relevant spans in unstructured privacy policies and the selection of the best-suited NLP model for this specific task. The initial results of a user evaluation are presented, showing promising directions. Eventually, some future research directions for the extension of the system conclude our contribution.

Keywords: Privacy · Personally Identifiable Information · Policy Comprehension · Websites' Privacy Regulation · Sentence Boundary Detection · Question to Document Matching · Natural Language Processing · Question Answering

A. Moallem (Ed.): HCII 2023, LNCS 14045, pp. 194–212, 2023.
https://doi.org/10.1007/978-3-031-35822-7_14

1 Introduction

Online privacy and data protection is a trending topic, both in research and within the political agenda [27]. In fact, many countries or geographical regions enforced regulations [8] to oversee the scope and the rights of this personal data collection and usage by companies. These privacy laws provide a significant level of privacy guarantees to people [33] and obligations on brands that process people's personal data [29].

The regulations and their enforcement have to deal with balancing between the level of protection for individuals' privacy and the legitimate and necessary usage of data as part of the information age (e.g. as protection under the Freedom of Information [1], which is generally guaranteed in the constitution of liberal countries). This is even more relevant in sectors such as social care [2] and health, where the quality of the cures and the advancements in medicine can be directly affected by the possibility to collect, manipulate and interpret medically-relevant parts of patients' personal data [13], at different levels of aggregation [28].

The problem is even more pervasive and hard to control in an online setting, where tracking technologies and information collection tools can be seamlessly embedded into web browsers and apps. In fact, everyone is affected by the phenomenon of the usage of personal data by online companies running websites and online services. As a demonstration, you are constantly asked to accept agreements to be able to access the information or application requiring you to give your "consent" to the processing of your personal information, even though you do not exactly know what you are consenting to [32].

Despite the fact that strict new regulations have been put into place in Europe (starting with the General Data Protection Regulation or GDPR) and other regions in the world, which protect the collection and use of customer's personal data, the biggest obstacle to their effectiveness remains people's inability to understand their legal rights and the lack of transparency from companies collecting data [4].

In our research project [15], we aim at supporting customers to understand practically the terms of a website's privacy policy before accepting it. In that direction, we are proposing a system that can identify relevant parts of an official website privacy policy, based on users' queries formulated in natural language. Instead of blindly accepting a privacy policy, a website user could first get a response to a concern he/she might have (e.g., "*I don't want to be targeted by email after reading an article on your site. Can you please confirm that I will not receive any marketing or promotional emails after I accept the privacy policy?*").

This is a first step towards better awareness and a higher comprehension rate regarding the permitted usage of the collected personal data by companies, and how customers can more effectively defend themselves whenever the terms and conditions are not fully respected [12].

2 Relevant Works

In this section, we present the main relevant works for the different subcomponents of our solution. First, we shortly introduce approaches for *Sentence Boundary Detection* followed by solutions for *Question to Document matching.*

2.1 Sentence Boundary Detection

Proposed by [11], *NLTK Punkt Tokenizer* is an unsupervised model that relies on the identification of abbreviations in a sentence. The authors argue that abbreviations can disambiguate sentence boundaries as the assumption is that an abbreviation is a collocation of the truncated word and its period. This collocational system has also shown efficiency in detecting initial and ordinal numbers. The method is very straightforward as it only needs the sentence itself and is not dependent on the context or language, an ideal feature when applied in a multilingual setting.

[22] proposed a rule-based sentence boundary disambiguation toolkit, *PySBD*, that has both universal rules shared across languages and language-specific rules. These rules for segmentation go from common rules (i.e. identification of main sentence boundaries, periods, single/multi-digit numbers, parentheses, time periods, etc.) to rules that handle geolocation references, abbreviations, exclamations, etc.

Another toolkit proposed by [19], *Stanza*, offers a fully neural pipeline for natural language processing (NLP) including tokenization, lemmatization, named entity recognition, and more. The tokenization model, in particular, combines tokenization and sentence segmentation by treating text as a tagging problem and predicting if a given character is the end of a word, a sentence or a multi-word token.

In the legal domain, [23] examined several models as legal text presents problems in terms of punctuation, structure and syntax, that common language does not have. Three models were considered: NLTK Punkt Tokenizer, Conditional Random Field (CRF) [14], and a neural network such as Word2Vec [16]. The author observed that a simple model such as NLTK Punkt Tokenizer might be a good choice in general but needs further training to give acceptable results in the legal domain. The best performance was given by the CRF approach since it resulted to be the most practical and simple model to train. As for the neural network, the author suggested to use more sophisticated word embeddings such as BERT [3] to obtain better and competitive results.

A legal dataset was created by [25] to help NLP models to segment US decisions into sentences. The dataset has sentence boundaries annotations made by human annotators and is composed by 80 US court decisions from four different domains resulting in more than 26000 annotations.

2.2 Question to Document Matching

IDF-Based. Usually adopted as a baseline for Question to Documents (Q2D) matching, *BM25-Okapi* [21] is an Inverse Document Frequency-based (IDF)

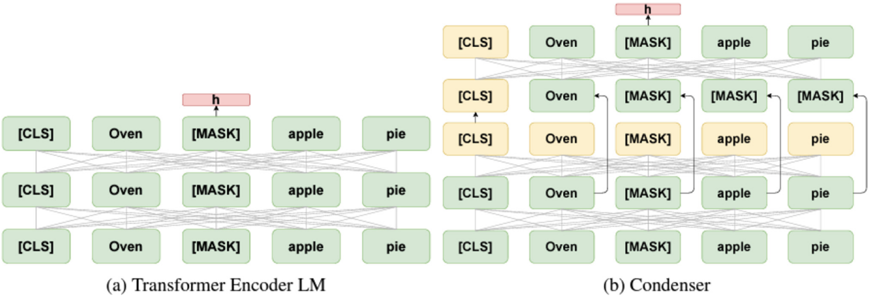


Fig. 1. A general Transformer LM architecture (a) vs. Condenser architecture (b) [7].

model that relies on rare words to match a query with documents by ranking their relevance. It is a computationally lightweight method reported in many scientific works, as in some cases it can still outperform heavier deep learning models. In addition to BM25-Okapi, several other variants of the BM25 algorithm have been proposed such as BM25-L and BM25+ [31].

Keyword-Based. Proposed by [26], *KeyBERT*[1] is a method for extracting keywords and keyphrases and find similarities between a sentence and a given document. It uses BERT word embeddings to extract document and sentence representations paired with cosine similarity to get the most similar documents to a given sentence. It is a quick, simple but powerful method that can be considered state-of-the-art in the keyword extraction domain.

Bi-encoders. The idea is to use pre-trained Transformer language models to extract the representations from queries and documents in an independent manner and compute their similarity with the dot product. However, pre-trained models, such as BERT, are not specifically trained to do retrieval out of the box so what most of the bi-encoder models try to do is fine-tuning. Furthermore, pre-trained models do not have an attention structure ready for bi-encoders, that is, they are not capable of aggregating complex data into single dense representations. In this regard, [6,7] argue that bi-encoder fine-tuning is not efficient as pre-trained models lack *structural readiness*. Thus, they proposed *Condenser*[2], a novel pre-training architecture that not only tries to fine-tune towards a retrieval task but, more importantly, is pre-trained towards the bi-encoder structure by generating dense representations (Fig. 1).

Cross-Encoders. As opposed to bi-encoders, cross-encoders compute the score between a query and documents by encoding them together. This enables, when using Transformers, full self-attention between queries and documents. However,

[1] https://maartengr.github.io/KeyBERT/index.html.
[2] https://github.com/luyug/Condenser.

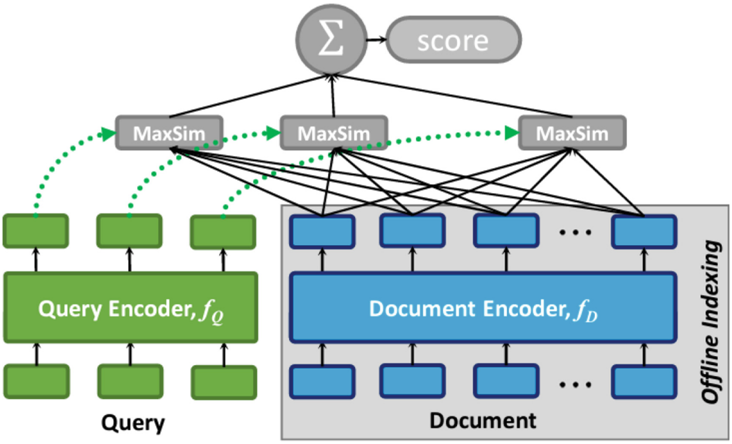


Fig. 2. Architecture of ColBERT given a query and a document [10].

such a powerful structure requires significant computational power as it has to do a forward pass through the model to obtain a score for each document. To reduce the computational burden, cross-encoders are usually combined with re-ranking. [17] proposed a cross-encoder combined with *BM25*[3] to narrow the searching space. Firstly, they retrieve a fixed number of relevant documents to a given query by using *BM25*. Secondly, they re-rank the retrieved documents by using BERT as a binary classification model. Finally, the top-k documents will be chosen as the candidate answers.

Hybrids. Hybrid architectures can be considered as a composition of bi-encoders and cross-encoders. Some models such as *ColBERT* [10][4], introduce a new ranking method, *late interaction*, to adapt language models, such as BERT, for retrieval (Fig. 2). The model encodes independently query and documents using BERT, re-ranks documents offline through pre-computation and computes the relevance between query and documents via late interaction that the authors define as a summation of maximum similarity. Santhanam et al. [24] then proceeded to enhance the model by producing *ColBERTv2*. It consists of the same architecture as *ColBERT* but with advances in quality and space efficiency of vector representations. This method is state-of-the-art.

Another method, *LaPraDoR*[5], proposed by [34], uses an unsupervised dual-tower model for zero-shot text retrieval that iteratively trains query and document encoders with a cache mechanism. Unlike supervised methods, this model combines lexical matching with semantic matching, achieving state-of-the-art results. Our own investigations of transformer model performances in the pri-

[3] https://github.com/nyu-dl/dl4marco-bert.
[4] https://github.com/stanford-futuredata/ColBERT.
[5] https://github.com/JetRunner/LaPraDoR.

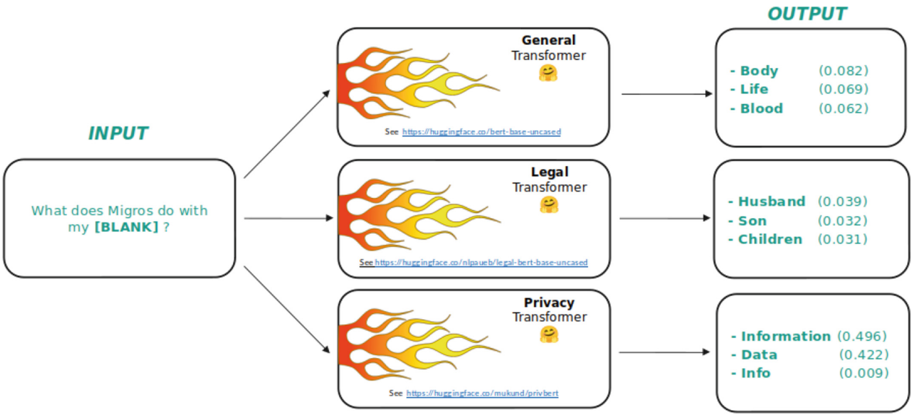


Fig. 3. A simple example of the effect of a domain-specific corpus in the training/fine-tuning of deep learning models. The same input query is matched with different words. This is explained by the different frequencies of co-occurrence in the specific realms.

vacy text domain are summarized in Fig. 3. We show that using a domain-specific corpus for training and/or fine-tuning of deep learning models leads to increased performance, thus justifying the need for a specialized model in privacy policy comprehension tasks.

3 Technical Solution

The design of the current demonstrator was based on recent approaches for serving deep learning (DL) models on the web. Figure 4 present the three-layer architecture orchestrated by *docker-compose* which also manages efficiently all the dependencies. The first layer (back-end supporting services) is composed of three parts.

1. A performant, flexible and easy to use tool for serving Machine (ML) models, called TorchServe[6]: here different DL models are served in a RESTful way. In particular, we plan to embed there the following models: BERT, SBERT, PrivBERT.
2. A vector database QDrant[7] able to store all the vector representations of the sentences and documents. This allows providing real-time answers to users, without the need to recompute the documents and sentence embeddings for every request.
3. A DBMS to store information, such as the TF-IDF representations of the documents.

[6] https://pytorch.org/serve/.

[7] Vector Search Engine QDrant, see https://qdrant.tech/.

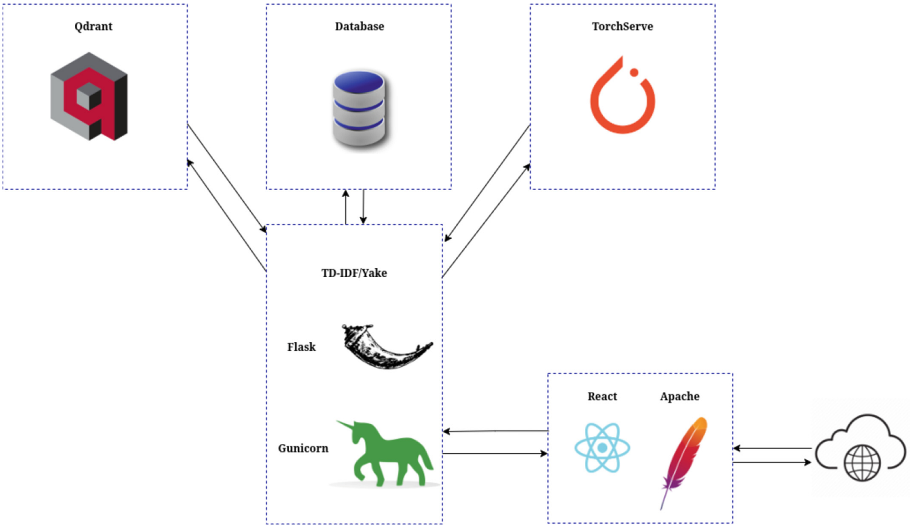


Fig. 4. The architecture of the solution in development. Everything is implemented as a multi-container Docker application, thus the orchestration and dependencies can be managed effectively.

The second layer is the core of the service, composed of a python-based RESTful interface relying on the library flask, Gunicorn, and Yake, while the third layer is the frontend, implemented as a web-based interface using the Apache2 web server and the React JavaScript library.

In the following subsections, we present two main technical aspects affecting the quality of results from our initial demonstrator. On one side, the identification of spans representing valid sentences, as the basic building blocks for the matching and, on the other side, the matching approach between the user query and the documents existing in our library.

3.1 Sentence Boundary Detection

To benchmark SBD for our project, we first proceed to find annotated SBD datasets which may be relevant to our case. One relevant dataset was proposed by [25] and consists of annotated sentence boundaries for legal US documents (hereby referred to as *Legal*). We find this useful for us since privacy documents could be considered as special legal documents. To construct another dataset, we sample 10 privacy policies crawled from [15] and perform SBD annotation on these policies. For this, we utilize five independent annotators who are familiar with privacy policies and conduct specialized annotation using the Label Studio community edition software [30]. We gather all annotations and resolve annotator conflicts using the majority decision. This produces a dataset hereby referred to as *Annotation*, where the Inter-Annotator Fleiss κ metric [5] is 0.707.

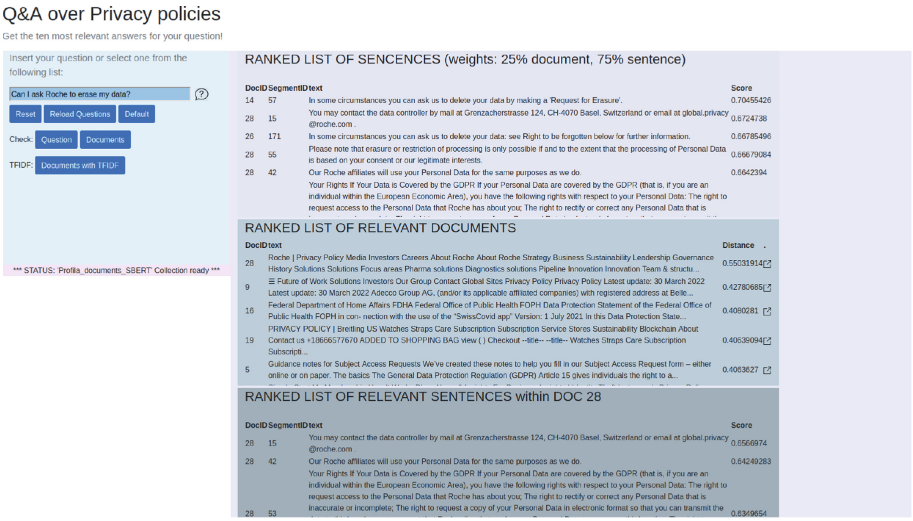
Table 1. Summary of SBD tokenizers, datasets, performances and runtime per sentence (in milliseconds)

Model	Dataset	Macro-F_1	Runtime (ms)
`nltk`	Legal	0.729	0.014
	Annotation	0.867	0.014
`pysbd`	Legal	0.656	1.571
	Annotation	0.689	0.944
`spacy`	Legal	0.682	1.894
	Annotation	0.681	2.189
`stanza`	Legal	**0.927**	3.221
	Annotation	**0.938**	4.606

With the legal and annotation SBD data sets, we proceeded to choosing competitive sentence tokenizers to benchmark. For this, we select the NLTK Punkt, PySBD, SpaCy and Stanza sentence tokenizers, hereby referred to as

Table 2. Tabular results of Q2D with models, datasets MRR@N metrics, precompute runtime per document (in milliseconds) and search runtime per query (in milliseconds).

Model	Dataset	MRR@1	MRR@5	MRR@10	Precompute runtime (ms)	Search runtime (ms)
TF-IDF	PrivacyQA	0.068	0.082	0.089	2.433	100.165
	Profila	0.577	0.639	0.647	2.433	93.153
BM25-L	PrivacyQA	0.047	0.055	0.062	2.345	0.509
	Profila	0.471	0.572	0.585	2.345	0.547
BM25-Okapi	PrivacyQA	0.079	0.097	0.106	2.297	0.486
	Profila	0.654	0.714	0.720	2.297	0.516
BM25+	PrivacyQA	0.076	0.093	0.103	2.354	0.496
	Profila	0.692	0.740	0.747	2.354	0.525
Db-Tas	PrivacyQA	0.079	0.108	0.120	323.305	17.358
	Profila	**0.712**	**0.762**	**0.766**	323.305	17.834
Db-Dot	PrivacyQA	0.074	0.101	0.113	309.471	17.183
	Profila	0.538	0.639	0.651	309.471	16.239
Rb-Ance	PrivacyQA	**0.089**	**0.114**	**0.123**	546.367	21.224
	Profila	0.615	0.688	0.695	546.367	21.972
ML-4	PrivacyQA	0.084	0.109	0.117	19.775	255.300
	Profila	0.673	0.746	0.756	19.775	265.205
ML-6	PrivacyQA	0.082	0.105	0.114	20.041	357.514
	Profila	0.654	0.728	0.739	20.041	372.114
ML-12	PrivacyQA	0.085	0.107	0.117	21.039	663.972
	Profila	0.663	0.744	0.752	21.039	690.180



Q&A over Privacy policies

Get the ten most relevant answers for your question!

Insert your question or select one from the following list:

Can I ask Roche to erase my data?

*** STATUS: 'Profila_documents_SBERT' Collection ready ***

RANKED LIST OF SENCENCES (weights: 25% document, 75% sentence)

DocID	SegmentID	text	Score
14	57	In some circumstances you can ask us to delete your data by making a 'Request for Erasure'.	0.70455426
28	15	You may contact the data controller by mail at Grenzacherstrasse 124, CH-4070 Basel, Switzerland or email at global.privacy @roche.com .	0.6724738
26	171	In some circumstances you can ask us to delete your data: see Right to be forgotten below for further information.	0.66785496
28	55	Please note that erasure or restriction of processing is only possible if and to the extent that the processing of Personal Data is based on your consent or our legitimate interests.	0.66679084
28	42	Our Roche affiliates will use your Personal Data for the same purposes as we do.	0.6642394
		Your Rights If Your Data is Covered by the GDPR If your Personal Data are covered by the GDPR (that is, if you are an individual within the European Economic Area), you have the following rights with respect to your Personal Data: The right to request access to the Personal Data that Roche has about you; The right to rectify or correct any Personal Data that is	

RANKED LIST OF RELEVANT DOCUMENTS

DocID	text	Distance	.
28	Roche | Privacy Policy Media Investors Careers About Roche About Roche Strategy Business Sustainability Leadership Governance History Solutions Solutions Focus areas Pharma solutions Diagnostics solutions Pipeline Innovation Innovation Team & structu...	0.55031914	↗
9	☰ Future of Work Solutions Investors Our Group Contact Global Sites Privacy Policy Privacy Policy Latest update: 30 March 2022 Latest update: 30 March 2022 Adecco Group AG, (and/or its applicable affiliated companies) with registered address at Belle...	0.42780685	↗
16	Federal Department of Home Affairs FDHA Federal Office of Public Health FOPH Data Protection Statement of the Federal Office of Public Health FOPH in con- nection with the use of the "SwissCovid app" Version: 1 July 2021 In this Data Protection State...	0.4080281	↗
19	PRIVACY POLICY | Breitling US Watches Straps Care Subscription Subscription Service Stores Sustainability Blockchain About Contact us +18666577670 ADDED TO SHOPPING BAG view ( ) Checkout --title-- --title-- Watches Straps Care Subscription Subscripti...	0.40639094	↗
5	Guidance notes for Subject Access Requests We've created these notes to help you fill in our Subject Access Request form – either online or on paper. The basics The General Data Protection Regulation (GDPR) Article 15 gives individuals the right to a...	0.4063627	↗

RANKED LIST OF RELEVANT SENTENCES within DOC 28

DocID	SegmentID	text	Score
28	15	You may contact the data controller by mail at Grenzacherstrasse 124, CH-4070 Basel, Switzerland or email at global.privacy @roche.com .	0.6566974
28	42	Our Roche affiliates will use your Personal Data for the same purposes as we do.	0.64249283
28	53	Your Rights If Your Data is Covered by the GDPR If your Personal Data are covered by the GDPR (that is, if you are an individual within the European Economic Area), you have the following rights with respect to your Personal Data: The right to request access to the Personal Data that Roche has about you; The right to rectify or correct any Personal Data that is inaccurate or incomplete; The right to request a copy of your Personal Data in electronic format so that you can transmit the	0.6349654

Fig. 5. The current prototype that implements the Architecture presented in Fig. 4.

nltk, pysbd, spacy and stanza respectively. The nltk, pysbd and stanza sentence tokenizers have been described in Sect. 2.1. spacy [9] is an additional sentence tokenizer which works by segmenting sentences using a dependency parser. Table 1 provides a summary of results from our SBD benchmarking process. To calculate the Macro-F_1 metric, we use a similar BIL character-token framework as per [23] and only use the statistic from the B and L character tokens, so as to prevent over-representation from I tokens. Our results show that the stanza sentence tokenizer outperforms all other tokenizers by a margin between 5% and 20% F_1 score. Additionally to Table 1, we provide visualizations of the results in Appendix A.

3.2 Question to Document Matching

The next pertinent technical problem in our project is finding relevant documents for each query. We refer to this problem as Q2D or Question to Documents. This is a well-known problem in NLP and falls under the general domain of Information Retrieval (IR), as described in Sect. 2.2. To benchmark Q2D, we start off by selecting appropriate datasets. We use *PrivacyQA* [20] and convert the dataset into a Q2D format, since its original format was designed for query-to-sentence tasks. Next, we select annotated data from [15] for Q2D and refer to this as *Profila*.

Based on Sect. 2.2, we select the following sparse Q2D models: TF-IDF, BM25-L, BM25-Okapi, BM25+ [31]. For dense models, we utilize bi-encoders and cross-encoders. The bi-encoders are Db-Tas, Db-Dot and Rb-Ance with the following Huggingface tags: sentence-transformers/msmarco-distilbert-base-tas-b, sentence-transformers/msmarco-distilbert-base-dot-prod-

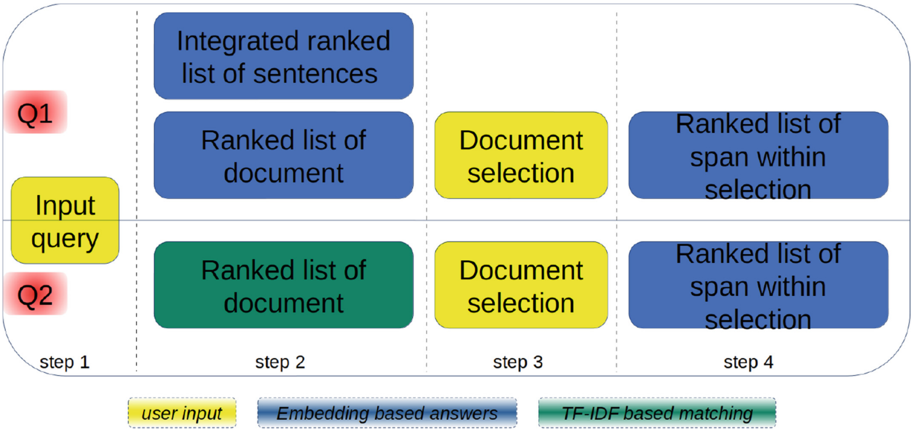
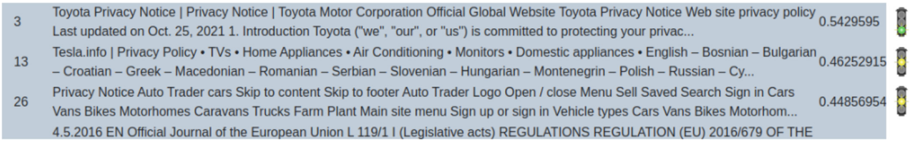


Fig. 6. The two pathways envisioned for the interaction with the GUI: the upper one is purely based on DL embedding, while the other uses the TF-IDF approach, as a first initial to match relevant documents.

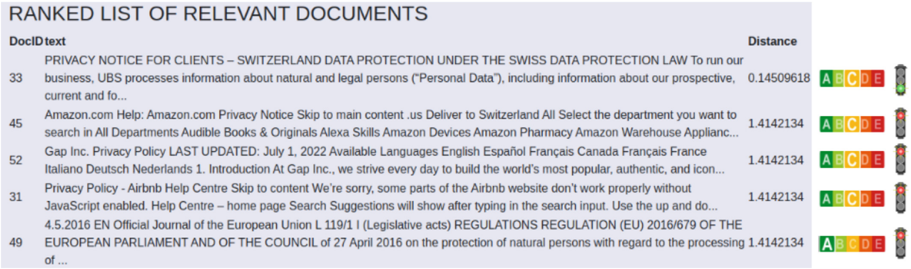
v3 and `sentence-transformers/msmarco-roberta-base-ance-firstp`. Cross-encoders consist of a BM25+ layer which minimizes the search space to the top 100 documents. These top documents are then fed into the cross-encoder to re-rank. The selected cross-encoders are ML-4, ML-6 and ML-12 which correspond to the following huggingface tags: `cross-encoder/ms-marco-MiniLM-L-4-v2`, `cross-encoder/ms-marco-MiniLM-L-6-v2` and `cross-encoder/ ms-marco-MiniLM-L-12-v2`.

We report the results of the Q2D benchmark in Table 2. We utilize the Mean Reciprocal Rank (MRR) metric with a cutoff for the top K documents. We utilize cutoffs of 1, 5, and 10 and, therefore, report the MRR@1, MRR@5 and MRR@10 metrics. We observe Db-Tas performs the best overall on the Profila dataset. Correspondingly, we observe Rb-Ance performs the best in the PrivacyQA dataset. Additionally to the Table 2, we visualize these results in Appendix A.



3	Toyota Privacy Notice | Privacy Notice | Toyota Motor Corporation Official Global Website Toyota Privacy Notice Web site privacy policy Last updated on Oct. 25, 2021 1. Introduction Toyota ("we", "our", or "us") is committed to protecting your privac...	0.5429595	🚦
13	Tesla.info | Privacy Policy • TVs • Home Appliances • Air Conditioning • Monitors • Domestic appliances • English – Bosnian – Bulgarian – Croatian – Greek – Macedonian – Romanian – Serbian – Slovenian – Hungarian – Montenegrin – Polish – Russian – Cy...	0.46252915	🚦
26	Privacy Notice Auto Trader cars Skip to content Skip to footer Auto Trader Logo Open / close Menu Sell Saved Search Sign in Cars Vans Bikes Motorhomes Caravans Trucks Farm Plant Main site menu Sign up or sign in Vehicle types Cars Vans Bikes Motorhom...	0.44856954	🚦
	4.5.2016 EN Official Journal of the European Union L 119/1 I (Legislative acts) REGULATIONS REGULATION (EU) 2016/679 OF THE		

Fig. 7. A mockup that adopts the semaphore metaphor to represent the match level between the requested query and the presented documents.



RANKED LIST OF RELEVANT DOCUMENTS

DocID	text	Distance
33	PRIVACY NOTICE FOR CLIENTS – SWITZERLAND DATA PROTECTION UNDER THE SWISS DATA PROTECTION LAW To run our business, UBS processes information about natural and legal persons ("Personal Data"), including information about our prospective, current and fo...	0.14509618
45	Amazon.com Help: Amazon.com Privacy Notice Skip to main content .us Deliver to Switzerland All Select the department you want to search in All Departments Audible Books & Originals Alexa Skills Amazon Devices Amazon Pharmacy Amazon Warehouse Applianc...	1.4142134
52	Gap Inc. Privacy Policy LAST UPDATED: July 1, 2022 Available Languages English Español Français Canada Français France Italiano Deutsch Nederlands 1. Introduction At Gap Inc., we strive every day to build the world's most popular, authentic, and icon...	1.4142134
31	Privacy Policy - Airbnb Help Centre Skip to content We're sorry, some parts of the Airbnb website don't work properly without JavaScript enabled. Help Centre – home page Search Suggestions will show after typing in the search input. Use the up and do...	1.4142134
49	4.5.2016 EN Official Journal of the European Union L 119/1 I (Legislative acts) REGULATIONS REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of ...	1.4142134

Fig. 8. Another proposal for the representation of the trustworthiness and authoritative level of each reported resource.

4 User Evaluation

In order to have an initial feedback on the current prototype we designed and ran an online survey, with a restricted set of potential users. In the survey we check different aspects of the prototype such as the quality of the proposed query to document matches, and the proposed design prototypes.

4.1 Questionnaire Design

The questionnaire is composed of 3 different parts. The first one is about the perceived ease of interaction with the demonstrator (see Fig. 5), in particular with respect to the two different pathways envisioned (see Fig. 6) namely the pure Deep Learning and the TF-IDF pathway. The second part is about the usage of graphical scales to report the trustworthiness (see Fig. 7) and the relevance of the match (see Fig. 8). The third one is about the next steps in the project: first, the type of information that seems to be relevant and important for creating the expert profile (see Fig. 10), and second, a different organization of information in the GUI, that seamlessly embed also the expert advice (see Fig. 9).

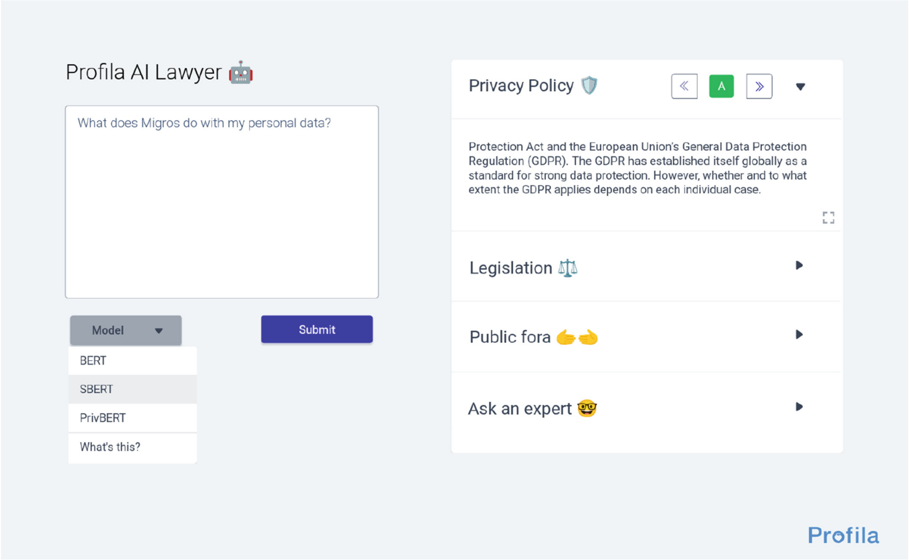
In Fig. 7, the semaphore metaphor is used to represent the relevance of the documents with respect to the query. The scale is dynamically applied to show groups with comparable relevance levels. The top group (in this case a single resource) is marked as green, while the next group is yellow and all the remaining matches are associated with a red semaphore, indicating that they are less relevant. An alternative approach we would like to explore could be to assume the score follows a standard distribution, and then compute the mean m and the standard deviation σ of the relevance score on the top-k resources. Thus, green could be assigned to resources with a value larger than $m + 2 * \sigma$ and red to resources with a score lower than $m - 2 * \sigma$ while all the other ones will be marked as yellow.

Table 3. Survey responses: quantitative (top) and qualitative part (bottom)

Code		Question	Scale	Mean	Std	Ref
Q1		Please, rate the intuitiveness of the service (pure embeddings)	1..5	2.82	1.17	Figure 6, top
Q2		Please, rate the intuitiveness of the service (TF-IDF + embeddings)	1..5	2.91	1.22	Figure 6, bottom
Q3	1	The semaphore metaphor is self-explanatory	1..10	7.36	1.45	Figure 7
	2	I prefer the semaphore over the numerical value	1..10	8.09	1.73	
	3	For me, it is easier to grasp the ranking using ONLY the semaphore icon	1..10	6.73	1.29	
	4	I would prefer the semaphore icon AND the numerical value	1..10	5.18	1.20	
	5	I would like a solution with the semaphore icon and the numerical value, ON MOUSE-OVER	1..10	6.45	1.10	
	6	I would like a solution with the semaphore icon and the numerical value, ON CLICK	1..10	3.82	1.45	
Q4	1	The Nutri-Score metaphor is self-explanatory	1..5	3.27	0.84	Figure 8
	2	I find the interface with the two icons overwhelming	1..5	3.27	1.10	
	3	I would prefer a scale with only 3 values (trustable - partially trustable - user-generated/doubt)	1..5	3.91	2.86	
	4	I would simply use a color scale, without a letter	1..5	3.91	2.81	
	5	I find this information valuable	1..5	3.27	1.30	
Q6		How would you rate this proposal	1..5	4.6	0.7	Figure 9

Q5 (**ref:** Figure 10)	**Relevance**			**Importance**			
	Not	*Partial*	*Very*	*Not*	*Partial*	*Somehow*	*Very*
document edited	0%	45%	55%	0%	27%	18%	55%
document contributed to	0%	36%	64%	0%	9%	27%	64%
answer provided	0%	9%	91%	0%	9%	9%	82%
activity on the platform	36%	55%	9%	9%	27%	45%	18%
self-declared knowledge	9%	82%	9%	18%	45%	18%	18%

Another proposition for the representation of the trustworthiness and authoritative level of each reported resource is presented in Fig. 8. This builds on top of the presented semaphore metaphor from Fig. 7. The scale work as follows: Dark Green (A) means those are (national or international) laws, where *Light Green (B)* matches with regulations, court, administrative cases, and privacy-oriented associations recommendations, with *Yellow (C)* official privacy policies or law-regulated agreements from institutions/companies are identified, while the two final categories Light (D) and Dark Red (E) indicate the resources that are user-generated (UGC) or found online on not-vetted resources, such as in public fora or non-professional new groups about legal and privacy issues.



What does Migros do with my personal data?

Protection Act and the European Union's General Data Protection Regulation (GDPR). The GDPR has established itself globally as a standard for strong data protection. However, whether and to what extent the GDPR applies depends on each individual case.

Fig. 9. A mockup of the potential Web-based GUI for the initial release of the "*Profila AI Lawyer*" service. Here, the user is guided by the responses' headers to understand the trustworthiness and authoritativeness level of each proposed resource.

4.2 Data Analysis

We collected 16 valid responses in the time span of a week from individual participants. Their profiles are heterogeneous, as they cover multiple roles and responsibilities within members of the project team, but also marketing, communication, and product engineering on the company side. A limited number of potential users were also included.

Table 3 presents a synoptic view over this initial survey. The first two questions (current demonstrator intuitiveness) show average values with significant variability, thus demonstrating the need for improvement in the way of presenting information and in the proposed interaction pathways. The third question, dealing with the semaphore metaphor, explores it in contrast to the current similarity values. The participants rated positively the intuitiveness of this analogy as a replacement for the numeric value, with the possibility to reveal it using a mouse-over approach. This question also exposes the participants' preference for a simpler and minimalist interaction approach (Q3.4 and, particularly, Q3.6). In question Q4 we proposed to use the additional metaphor of the Nutri-Score as the information carrier [18] for the trustworthiness and authoritativeness level of each reported resource. Its intuitiveness is rated quite positively, even if its simplification to a limited set of three values (see Q4.2) represented by a single color without an alphabetical label could be even preferred by some users (even if with a significantly larger variability, see Q4.4). The use of the two icons is anyway almost constantly not perceived as overwhelming. Another aspect covered in the survey, even if in a purely qualitative way, is the sources relevant and

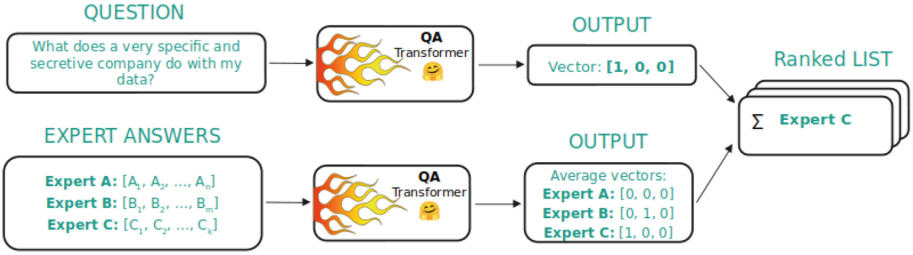


Fig. 10. The envisioned solution for matching the best-suited scholars (in terms of expertise and correct level of knowledge) to a privacy-oriented user query that did not receive a satisfactory answer through the Q&A self-help approach presented in Fig. 9.

important to be included in the experts' profile (see Table 3, bottom). Here it seems pretty evident that the document edited and contributed to, together with answers provided to customers' queries form the most relevant part. It is very important that these aspects are considered by future iterations of the platform, in order to obtain an accurate and reliable profiling process.

Other activities in the platform, including also the event of self-declaring skills, knowledge and/or competencies are perceived as less or not relevant and should be less or not at all included in the profiling process at run-time. Nevertheless, even if not perceived by the average user of this platform, this information can be relevant to solve the cold-start problem, where data about experts' contributions are very limited or absent. Eventually, the last question Q6 explores an alternative approach to display the trustworthiness and authoritativeness level of resources matching a user query, by grouping them into the categories of legislation, official privacy policies, and public fora/user-generated content. Additionally, the function to forward the customer request for support to one or more relevant legal scholars is presented as an additional option, then having a seamless integration into the remainder of the platform demonstrator. This mock-up was rated as very appealing by all the participants in the survey. We are planning to refine the questionnaire and extend its panel of participants, to obtain even more insights in the continuation of the project.

5 Conclusions and Future Work

Supporting consumers' comprehension of privacy policies and usage of their personal information collected online is an open problem. Legal agreements regulating this subject are usually difficult to interpret for the general public, due to their length and their domain-specific language and formulation.

This work presents a first prototype for an interface to extract relevant sections from privacy policies based on user queries in natural language. This contribution details the aims, the current status and the immediate future steps of a joint research project aimed at solving these issues by means of question answering within existing legislation and privacy policies, with the possibility to seamlessly obtain inexpensive professional punctual support for the more complex issues. In particular, the two aspects of *Sentence Boundary Detection* and of *Question to Document matching* were identified as particularly important for the quality of the provided results, and their effects were initially explored. To sum up, our main contributions detailed in this work are as follows.

1. We compare different SBD approaches specifically in the domain of privacy-related legal documents. The results demonstrate that the *stanza* sentence tokenizer delivers the best results in our use case clearly outperforming competing tools such as *nltk*, *pysbd* or *spacy*.
2. Our work features a technical evaluation of different automatic information retrieval models of different complexity, ranging from pure IDF- and keywords-based to bi-encoder and cross-encoder solutions, which indicates that a relatively light-weight and sparse IDF-based model (BM25+) practically outperforms other approaches when considering accuracy and efficiency aspects.
3. We present a user interface and architecture for delivering the results of the presented IR algorithms on privacy policy documents of potential customers.
4. We provide a user evaluation of the presented user interface which gives insights into the user comprehension of specific design decisions of our first prototype and sets a baseline to measure improvements of further iterations of the tool. This initial survey showed some promising results about the users' perception but also definitive areas of improvement that we need to tackle in order to make the service effective.

Based on these results and the general objective detailed, the list of next research steps is the following.

– We will realize the second part of the application, which will feature the transfer of queries to legal professionals based on multifaceted expert profiles (see Fig. 10). Here we will test different options, mainly based on the perception of relevance and importance of different user activities within the platform, as indicated by the survey results.
– Further experiments with the best-performing Q2D models will be carried out. One key point will be to explore why sparse lexical approaches outperform dense NN-based ones, and to use this information to reduce the complexity of the search while maintaining acceptable performances in the matching process.
– The user interface will be improved based on user evaluation. We will implement the mock-up (Fig. 9) of the next version presented in the user evaluation.

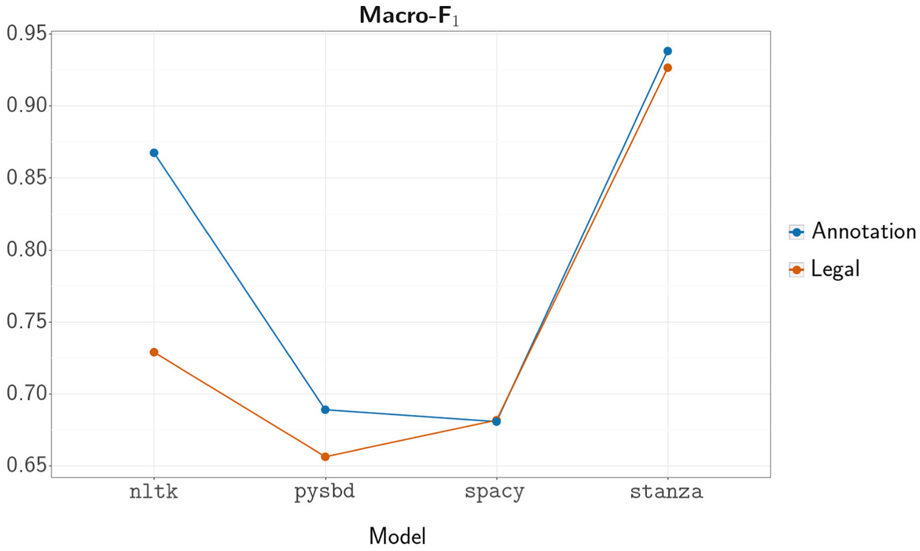
With these points, we aim at providing an effective solution to the presented problem, while advancing the state of the art in the area of domain-specific question answering for privacy policies and of heterogeneous profiling for similarity matches.

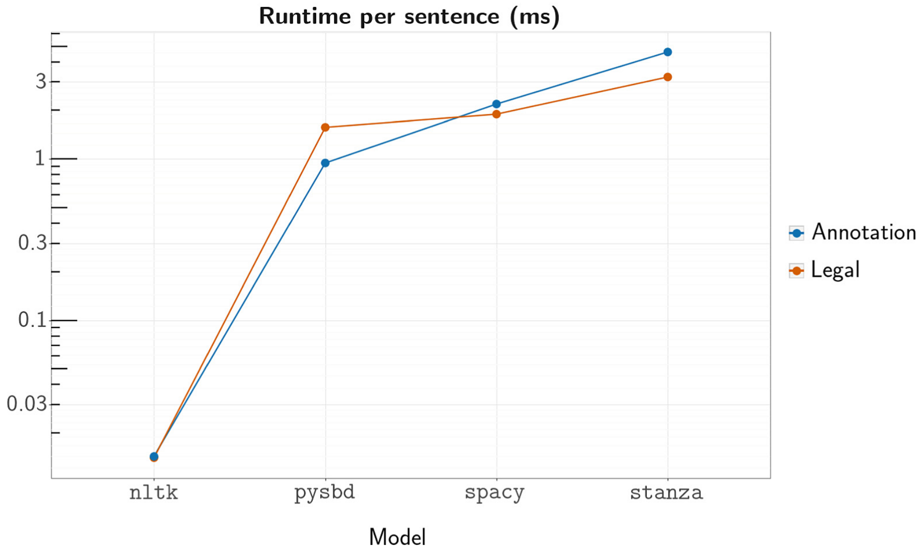
Acknowledgements. The research leading to this work was partially financed by *Innosuisse* - Swiss federal agency for Innovation, through a competitive call. The project 50446.1 IP-ICT is called *P2Sr Profila Privacy Simplified reloaded: Opensmart knowledge base on Swiss privacy policies and Swiss privacy legislation, simplifying consumers' access to legal knowledge and expertise* (https://www.aramis.admin.

ch/Grunddaten/?ProjectID=48867). The authors would like to thank all the people involved on the implementation side at Profila GmbH (https://www.profila.com/) for all the constructive and fruitful discussions and insights provided about privacy regulations and consumers' rights.

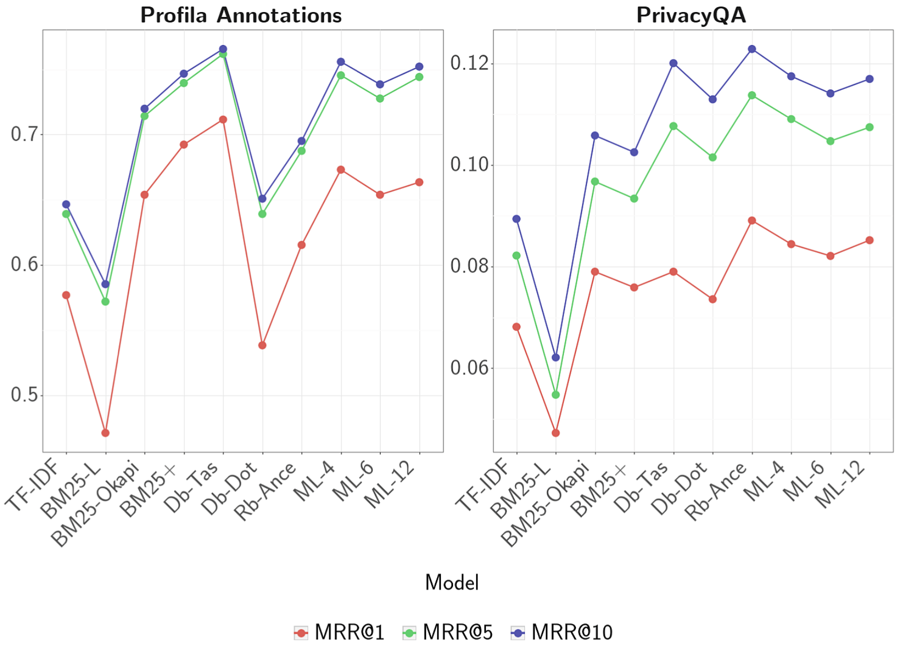
Appendix A - SBD and Q2D Graphs

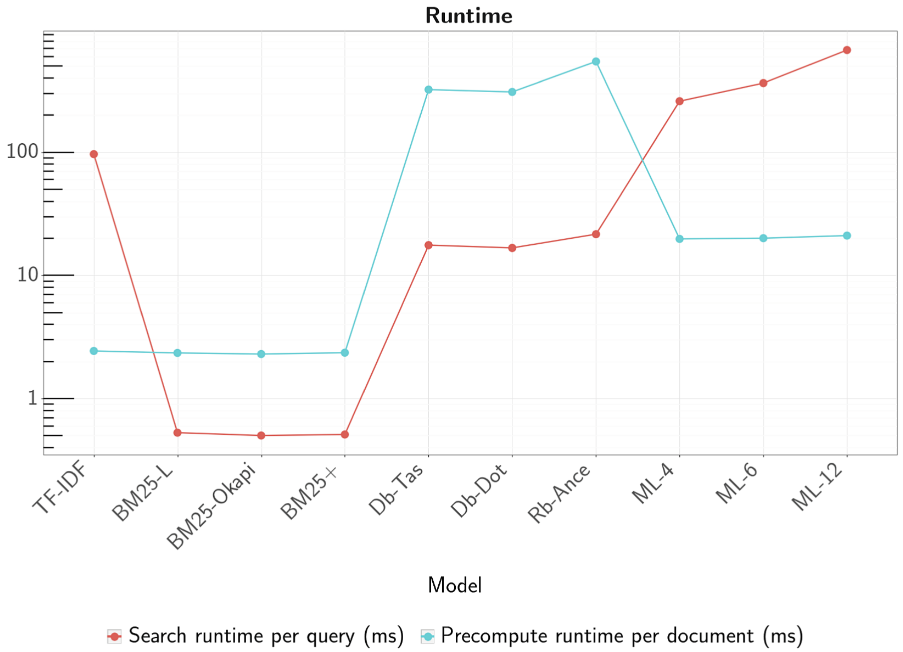
In this appendix, we provide the reader with the graphical representations of the data from Table 1 and from Table 2. Effectiveness of *nltk* is demonstrated with a good F1 measure and a very limited runtime.





BM25+, a relatively simple and sparse IDF-based model, practically outperforms other approaches when considering accuracy and runtime.





References

1. Abela, S.: Data protection and freedom of information. In: Abela, S. (ed.) Leadership and Management in Healthcare, pp. 103–107. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-21025-9_10
2. Crook, M.: The Caldicott report and patient confidentiality (2003)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Fabian, B., Ermakova, T., Lentz, T.: Large-scale readability analysis of privacy policies. In: Proceedings of the International Conference on Web Intelligence, pp. 18–25 (2017)
5. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378 (1971)
6. Gao, L., Callan, J.: Condenser: a pre-training architecture for dense retrieval. arXiv preprint arXiv:2104.08253 (2021)
7. Gao, L., Callan, J.: Is your language model ready for dense representation fine-tuning. arXiv preprint arXiv:2104.08253 (2021)
8. Goddard, M.: The EU general data protection regulation (GDPR): European regulation that has a global impact. Int. J. Mark. Res. **59**(6), 703–705 (2017)
9. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al.: Spacy: industrial-strength natural language processing in Python (2020)
10. Khattab, O., Zaharia, M.: ColBERT: efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–48 (2020)
11. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Comput. Linguist. **32**(4), 485–525 (2006)
12. Korunovska, J., Kamleitner, B., Spiekermann, S.: The challenges and impact of privacy policy comprehension. arXiv preprint arXiv:2005.08967 (2020)
13. Leatherman, S., Berwick, D.M.: Accelerating global improvements in health care quality. JAMA **324**(24), 2479–2480 (2020)
14. Liu, Y., Stolcke, A., Shriberg, E., Harper, M.: Using conditional random fields for sentence boundary detection in speech. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 451–458 (2005)
15. Mazzola, L., Waldis, A., Shankar, A., Argyris, D., Denzler, A., Van Roey, M.: Privacy and customer's education: NLP for information resources suggestions and expert finder systems. In: Moallem, A. (ed.) HCII 2022. LNCS, vol. 13333, pp. 62–77. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05563-8_5
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
17. Nogueira, R., Cho, K.: Passage re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019)
18. Peters, S., Verhagen, H.: An evaluation of the nutri-score system along the reasoning for scientific substantiation of health claims in the EU—a narrative review. Foods **11**(16), 2426 (2022)
19. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: a Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)

20. 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), Hong Kong, China, pp. 4949–4959. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1500. https://www.aclweb.org/anthology/D19-1500
21. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at trec-3. NIST Special Publication Sp **109**, 109 (1995)
22. Sadvilkar, N., Neumann, M.: PySBD: pragmatic sentence boundary disambiguation. arXiv preprint arXiv:2010.09657 (2020)
23. Sanchez, G.: Sentence boundary detection in legal text. In: Proceedings of the Natural Legal Language Processing Workshop 2019, Minneapolis, Minnesota, pp. 31–38. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-2204. https://aclanthology.org/W19-2204
24. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2112.01488 (2021)
25. Savelka, J., Walker, V.R., Grabmair, M., Ashley, K.D.: Sentence boundary detection in adjudicatory decisions in the United States. Traitement automatique des langues **58**, 21 (2017)
26. Sharma, P., Li, Y.: Self-supervised contextual keyword and keyphrase retrieval with self-labelling (2019). https://www.preprints.org/manuscript/201908.0073/v1
27. Sivan-Sevilla, I.: Varieties of enforcement strategies post-GDPR: a fuzzy-set qualitative comparative analysis (FSQCA) across data protection authorities. J. Eur. Public Policy 1–34 (2022)
28. Subrahmanya, S.V.G., et al.: The role of data science in healthcare advancements: applications, benefits, and future prospects. Irish J. Med. Sci. (1971-) **191**(4), 1473–1483 (2022)
29. Tikkinen-Piri, C., Rohunen, A., Markkula, J.: EU general data protection regulation: changes and implications for personal data collecting companies. Comput. Law Secur. Rev. **34**(1), 134–153 (2018)
30. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020–2022). Open source software https://github.com/heartexlabs/label-studio
31. Trotman, A., Puurula, A., Burgess, B.: Improvements to BM25 and language models examined. In: Proceedings of the 2014 Australasian Document Computing Symposium, pp. 58–65 (2014)
32. Vail, M.W., Earp, J.B., Antón, A.I.: An empirical study of consumer perceptions and comprehension of web site privacy policies. IEEE Trans. Eng. Manag. **55**(3), 442–454 (2008)
33. Vanberg, A.D.: Informational privacy post GDPR-end of the road or the start of a long journey? Int. J. Hum. Rights **25**(1), 52–78 (2021)
34. Xu, C., Guo, D., Duan, N., McAuley, J.: LaPraDoR: unsupervised pretrained dense retriever for zero-shot text retrieval. arXiv preprint arXiv:2203.06169 (2022)