KnIGHT: Mapping Privacy Policies to GDPR

Conference Paper · August 2018

CITATIONS
0

READS
186

3 authors, including:

Najmeh Mousavi
University of Bonn
8 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Simon Scerri
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
73 PUBLICATIONS **709** CITATIONS
SEE PROFILE

KnIGHT: Mapping Privacy Policies to GDPR

Najmeh Mousavi Nejad[1,2], Simon Scerri[1,2], and Jens Lehmann[1,2]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
{nejad, scerri, jens.lehmann}@cs.uni-bonn.de
http://sda.cs.uni-bonn.de
[2] Fraunhofer Intelligent Analysis and Information Systems (IAIS), Germany
https://www.iais.fraunhofer.de

Abstract. Although the use of apps and online services comes with accompanying privacy policies, a majority of end-users ignore them due to their length, complexity and unappealing presentation. In light of the, now enforced EU-wide, General Data Protection Regulation (GDPR) we present an automatic technique for mapping privacy policies excerpts to relevant GDPR articles so as to support average users in understanding their usage risks and rights as a data subject. *KnIGHT* (Know your rIGHTs), is a tool that finds candidate sentences in a privacy policy that are potentially related to specific articles in the GDPR. The approach employs semantic text matching in order to find the most appropriate GDPR paragraph, and to the best of our knowledge is one of the first automatic attempts of its kind applied to a company's policy. Our evaluation shows that on average between 70-90% of the tool's automatic mappings are at least partially correct, meaning that the tool can be used to significantly guide human comprehension. Following this result, in the future we will utilize domain-specific vocabularies to perform a deeper semantic analysis and improve the results further.

Keywords: Privacy policy · General Data Protection Regulation · Semantic text matching.

1 Introduction

As technologies for analysis of Web data have grown, data privacy concerns (e.g., fraud, identity theft, etc.) among end-users have become a major issue. Enterprises try to employ automated techniques to analyze customer's personal data in order to achieve their business goals, but unsurprisingly they do not always adhere to the law. In 2015, the Belgian privacy commission reported that Facebook's privacy policy breaches European law[3]. Moreover, in 2017 the Dutch data protection authority (DPA) announced that Microsoft's Windows 10 breaches data protection law, since it is not clear which personal data it collects

[3] https://www.theguardian.com/technology/2015/feb/23/facebooks-privacy-policy-breaches\%2Deuropean-law-report-finds

and why[4]. As a result, people apply self protection methods to ensure that their personal information are not being misused. According to a study, the users attempt to protect their data either by installing a privacy protection software or by providing false information to the websites [2]. In addition to the public's distrust, the majority also tend to skip the privacy policies when signing up for products and services. A survey claims that only 26% of participants in a lab study read privacy policies and this percentage is expected to be much lower outside of laboratory conditions [15]. Moreover, these findings were verified by another recent experiment in which 543 university students were asked to agree to a fictitious social network's privacy policy and terms of service in order to join the network [17]. According to the study only 26% did not select the 'quick join' and unsurprisingly the average reading time of those who read privacy policy was 73 seconds.

The supervisory authorities worldwide enact strict regulations regarding data protection of a natural person. The European Union as well, have recently upgraded the current data protection directive to harmonize data privacy laws across Europe. The (GDPR)[5] legislation came into force on May 25th, 2018. All EU member states must now comply with GDPR and other corresponding regulatory documents.

The novel approach behind *KnIGHT* exploits semantic similarity between words to associate the privacy policy sentences to the corresponding paragraphs in GDPR. We investigate text mining techniques that match privacy policy segments with relevant GDPR articles. The targeted beneficiaries of our tool are regular users who would like to become more aware of the contents of a privacy policy. *KnIGHT* offers them shortcuts to the underlying legislation so that they can learn more about their risks and rights; empowering them with the possibility to stop using a specific service if its privacy policy includes suspicious clauses, or to report it to an authority. Nevertheless, more advanced users (e.g. lawyers, legal experts and compliance officers) would also benefit from future improved versions of *KnIGHT*.

KnIGHT takes a privacy policy and GDPR articles and finds the correlations between the two documents at sentence and paragraph level. Since not all sentences in a privacy policy are related to data usage and analysis, first the candidate sentences will be identified using a *GATE* pipeline [9]. This step will significantly reduce the number of processed excerpts in the policy text. After that, for each candidate the most related article will be found. Finally, the best paragraph match from the identified article will be detected for that candidate sentence. It is worth to mention that *KnIGHT* is independent of policy type or the regulatory documents. Whenever a policy must comply with some laws, this technique can be applied. However, since the recently enforced GDPR has become of crucial importance it is the focus of this study.

[4] https://www.zdnet.com/article/microsofts-windows-10-breaches-data\%2Dprotection-law-say-dutch-regulator

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

To the best of our knowledge, in spite of the importance of this issue, there is no automated approach to match a company's policy to regulatory documents in order to assist regular end-users to check the lawfulness of a company's activities and to familiarize themselves with their rights as a consumer. As explained in section 2, although there have been numerous efforts addressing regulatory compliance, the objectives were somehow different. In section 3, we describe *KnIGHT*'s workflow and its architecture. In section 4 the experiments are explained and finally section 5 concludes this paper with a list of possible directions for future work.

2 Related Work

We compare and contrast our approach with efforts in three main categories: (1) regulatory compliance (2) semantic text matching and (3) privacy policy studies.

GaiusT is a tool which semi-automatically extracts rights and obligation from a regulatory document [20]. According to the paper, these obligations should be considered in the requirement model of a system software design. For extracting rights, duties and actors from a regulatory text, a semantic annotation framework is applied and extended. Finally, the approach is verified by presenting the result of two case studies: the U.S. Health Insurance Portability and Accountability Act (HIPPA) and the Italian accessibility law for information technology instruments. In a similar study, a process called "Semantic Parameterization" has been applied to privacy rules from HIPAA [5]. First, rights and obligations are reformulated into Restricted Natural Language Statements (RNLS) and then RNLSs are mapped into semantic models to clarify ambiguities. At the end, the authors listed some limitation which arose from their incomplete set of phrase heuristics. Last but not least, a prominent group in regulatory compliance is GRTCT[6]. The group has recently developed *Ganesha* platform which uses semantic technologies and assists the industry professionals to track and compare legislations, leading to better financial compliance. It employs OMG Standards[7] (SBVR, FIBO, FIRO, etc.) to semantically enrich the regulatory texts.

As opposed to our approach, none of the above efforts prioritize the regular end-users needs and concerns. Our goal is to design a policy reading system to assist consumers to familiarize themselves with a specific policy and the relevant regulation. We strive to motivate them to make themselves aware of what they are agreeing to by using a service and encourage them to know their rights as a citizen according to their regional laws. Furthermore, since the previous methods have dealt with a single target regulatory document, they are not suitable to be applied in our approach. Our identified problem tackles the overlap of a company's policy and the relevant legislations and therefore is addressing two sources of legal texts.

In the context of semantic text matching, the closest field to our approach is intelligent plagiarism. Intelligent plagiarism techniques find semantically similar

[6] `http://www.grctc.com/`
[7] `http://www.omg.org/`

matches in different documents (e.g., finding similar paragraphs in two papers). An extensive study of state-of-the-art has been conducted in [3]. The paper focuses on two main steps for plagiarism detection: first the latest techniques for retrieval of candidate documents are explained; then the exhaustive analysis of suspicious candidates for plagiarism detection is presented. For the first step two main approaches are usually used: information retrieval models (fingerprints, hash-based models, vector space models, etc.) and clustering techniques. For the second step, depending on the plagiarism type (literal or intelligent), different methods can be applied. In the case of intelligent plagiarism, semantic and fuzzy based method are used. Semantic features include word synonyms, antonyms, hypernyms and hyponyms. Furthermore, the use of thesaurus dictionaries and lexical databases gives more insights into the semantic meaning of the text. In addition, POS tagging and semantic dependencies will enrich semantic based methods. Regarding fuzzy based methods, word embeddings are similar to the "fuzzy" concept, since both implement a spectrum of similarity values for each word, e.g., there is a degree of similarity for each word and the associated fuzzy set. Finally, the authors recommend that semantic and fuzzy methods are the most proper approaches for intelligent plagiarism detection.

In contrast to our problem, intelligent plagiarism performs the information retrieval step only once to find the candidate documents, whereas in our case, the retrieval process for finding the related `GDPR` article(s) should be applied for each candidate sentence in the privacy policy. Therefore, we need a better solution to decrease the computation cost. However, we learned from the state-of-the-art that the most appropriate approach for finding similar matches is the semantic methods and hence we exploit semantic techniques in *KnIGHT*.

Some studies dealt specifically with privacy policies. Costante et al. and Guntamukkala et al. developed a system which evaluates the completeness of privacy policies [7, 11]. To achieve this goal, first some categories are defined following privacy regulations and guidelines. Then, employing text categorization and machine learning techniques, the corresponding categories of privacy policy paragraphs are determined. In this case, the user can examine the policy in a structured format and study those parts that she is interested in. Although both papers propose promising approaches, most of current privacy policies are already in a structured format and use rich HTML representation (for example *ReseachGate* [8] & *Unilever*[9]). In addition, both approaches have used supervised machine learning and complied a training data set manually whereas our method is independent of the subject material.

The Usable Privacy Policy Project initiated in 2013, aims to extract information from privacy policies and make it available for consumers, developers and regulators[10]. The team has recently launched an online service called Polisis[11], which exploits natural language processing and deep learning techniques to ex-

[8] `https://www.researchgate.net/privacy-policy`
[9] `http://www.unileverprivacypolicy.com/en_gb/policy.aspx`
[10] `https://usableprivacy.org/`
[11] `https://pribot.org/polisis`

tract segments from a privacy policy [12]. Each segment has a set of labels which describe its data practices. In order to train the model, the authors leveraged the *Online Privacy Policies* (OPP-115) dataset [19] and reached 88.4% accuracy for the automatic generation of labels. Despite the encouraging work done by the whole team, we believe that the missing part is relating the privacy policies to the data protection laws in the favor of both end-users and regulators. Nevertheless, analyzing their open annotated dataset (human & machine-generated) and utilizing the segment's labels for improving our mapping approach is one of our future direction (which was not feasible in this phase due to recent date of publication).

3 Approach

In this section we explain the architecture and implementation of *KnIGHT* which builds on *GATE* embedded and *Deeplearning4j* [1] open source APIs. *Deeplearning4j* or *DL4J* implements deep learning algorithms with a specific focus on neural network techniques. The library offers word2vec and paragraph2vec as well, with a default word2vec model trained on Google News Corpus[12]. Figure 1 shows the architecture and workflow of *KnIGHT* that consists of two main steps: preparation phase which is independent of input; and the main semantic matching phase. Each of the following subsections presents how each phase fits within the architecture.

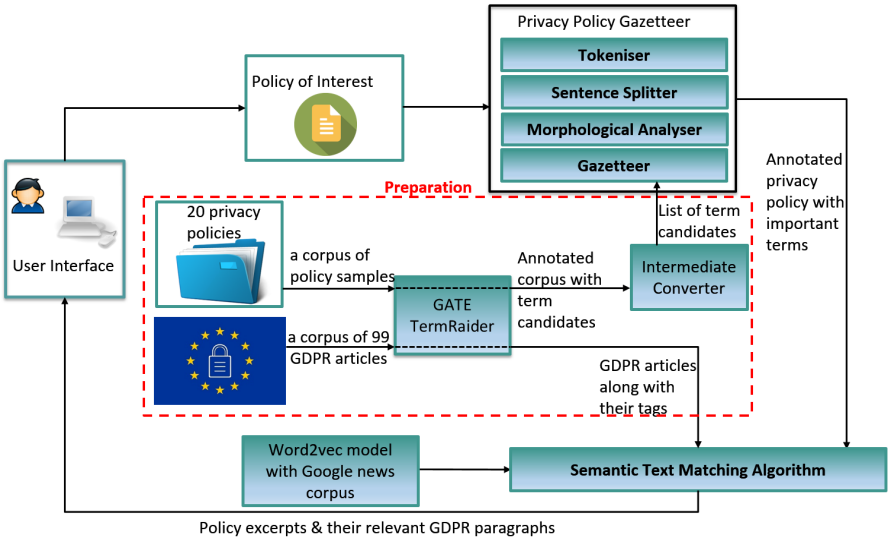

Fig. 1. Architecture and Workflow of *KnIGHT*

[12] https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors\%2Dnegative300.bin.gz

3.1 Preparation

Our approach deals specifically with GDPR legislation, therefore the pre-processing procedure can be done independently from the input (which is a privacy policy in the natural language). The preparation phase exploits a ready-made application called *GATE TermRaider*[13]. *TermRaider* is an English term extraction tool that runs over a text corpus and produces noun phrase term candidates together with a score which shows the salience of each term candidate in a domain specific corpus. Benefiting from this plugin, the preparation phase includes the following steps:

1. Twenty privacy policies from European Union companies were collected to build a privacy policy corpus.
2. Having this corpus, *TermRaider* was executed on top of it to find the most important terms in privacy policies which carry essential information. This step creates an annotation set called *Term Candidate*.
3. The annotation set produced in the previous step is converted to a text file to be used in the semantic text matching phase. Therefore, an intermediate converter processes all *Term Candidate* annotations and generates a list of terms with their corresponding roots (root is only meaningful when the term is a single token).
4. Another corpus was built with all 99 GDPR articles and *TermRaider* was executed on this corpus separately to generate a set of tags (also known as fingerprints) for each GDPR article. These tags are used in the related article retrieval phase (explained in the next subsection).

Since the preparation phase happens only once, the final response time will be reduced significantly. Furthermore, this layer architecture enables us to add more data privacy legislations in future with a small effort.

3.2 Semantic Text Matching

Once the initial processing has been done, the system will be ready to accept a privacy policy. As mentioned before, *KnIGHT* relates a sentence in a policy to (a) paragraph(s) in GDPR. The rational behind choosing the sentence level in the privacy policy is the existence of different layouts in writing those policies, e.g., it is complicated to determine the size and boundaries of a paragraph in an arbitrary policy. On the other hand, specifying the boundaries of a sentence is much easier in any form of a page style. However, processing all sentences in a privacy policy and relating them to GDPR is not logical, since some sentences carry service specific information and do not have a direct connection to GDPR, e.g., *Ryanair* says: "You will have the option to stay signed-in into your myRyanair account by checking the **remember me** box".[14] Processing these kind of sentences will only impose extra computation cost on the system without

[13] https://gate.ac.uk/projects/neon/termraider.html
[14] https://www.ryanair.com/gb/en/corporate/privacy-policy

any valuable result. Therefore, a simple pipeline called *Privacy Policy Gazetteer* will first find candidate sentences that have the potential to be matched to `GDPR` paragraphs.

Privacy Policy Gazetteer
We have created a pipeline using *GATE* embedded which contains some common preprocessing steps in NLP (tokeniser, sentence splitter, root finder) and a gazetteer that includes a list of important terms in a privacy policy. As described in subsection 3.1, the input text file for this gazetteer was compiled using *TermRaider* and an intermediate converter. A successful execution of this pipeline will create the following annotation types:

i) *Token* along with root feature;
ii) *Sentence*; and
iii) *Important Term*.

If a *Sentence* includes at least three *Important Terms*, it will be considered as a candidate.

Semantic Text Matching Algorithm
This component is the main element of *KnIGHT* and has three inputs: the annotated privacy policy with *Important Term* annotations, `GDPR` articles along with their tags and a word2vec model. Algorithm 1 shows the sketch of our semantic matching approach and has two main steps for each candidate sentence:

i) Retrieval of the most related `GDPR` article (line 3 to 14).
ii) Finding the best paragraph match in the identified article (line 15 to 25).

In the first step, the most related `GDPR` article is found for each candidate sentence. To achieve this goal, we compare the semantic similarity between two sets: Set_1 which contains important terms in the current candidate sentence and Set_2 that loops over all `GDPR` articles and in each loop, it contains the corresponding article tags.

Assuming sets S_1 and S_2 consist of n and m terms corresponding to T_{11}, ..., T_{n1} and T_{12}, ..., T_{m2}, the similarity between the two sets is calculated as shown in equation 1. In this equation $compositionalSim(Term1, Term2)$ is an extension of word2vec similarity function. Word2vec represents every word as an n-dimensional vector and then computes the semantic similarity between two words using *Cosine* similarity of two vectors. However, the default library does not provide a function for computing the similarity between multi-words terms. To solve this issue, we have defined a formula (equation 2) which composes all individual words vectors in a multi-words term by summation and creates a single vector for that term. Having two composed vectors for each multi-words term, the *Cosine* function is again applied to calculate the similarity between two terms. Finally, if the similarity between two sets is greater than a fixed threshold (line 9 in algorithm 1), it will be added to a list along with the similarity score. In the presence of a gold standard, the threshold can be determined by changing

Algorithm 1 Sketch of Text Matching Algorithm

Require: privacy policy candidate sentences, GDPR fingerprints, word2vec model
for all candidate sentences in the privacy policy **do**
  $candidateSentence \leftarrow$ current sentence
  $Set1 \leftarrow$ all important terms in $candidateSentence$
  $MatchesList \leftarrow$ an empty list
  for all GDPR articles **do**
    $ArticleNum \leftarrow$ article number
    $Set2 \leftarrow$ article tags
    $Sim \leftarrow$ similarity between $Set1$ & $Set2$
    if $Sim1 >$ threshold **then**
      add $Sim1$ & $ArticleNum$ to $MatchesList$
    end if
  end for
  $SortedList \leftarrow$ Sort $MatchesList$ acc. to sim
  $BestArticleMatch \leftarrow SortedList[0]$
  $MaxSim \leftarrow 0$
  $Vec1 \leftarrow$ word2vec vector of $candidateSentence$
  for all paragraphs in $BestArticleMatch$ **do**
    $currPar \leftarrow$ current paragraph
    $Vec2 \leftarrow$ word2vec vector of $currPar$
    $Sim2 \leftarrow$ similarity between $Vec1$ & $Vec2$
    if $Sim2 > MaxSim$ **then**
      $MaxSim \leftarrow Sim2$
      $bestParMatch \leftarrow$ current paragraph
    end if
  end for
end for
Ensure: Policy excerpts & their relevant GDPR paragraphs

its value and finding the best one which produces the highest F-measure. Due to lack of such gold standard in our case, the threshold is calculated by considering the average of all the similarity scores. Last but not least, our approach is able to find the TOP-n matches of GDPR. However for simplicity, only the best match is shown in the algorithm sketch.

$$Sim(S_1, S_2) = \frac{\sum_{i=1}^{n} \underset{1 \le j \le m}{max}[CompositionalSim(T_{i1}, T_{j2})]}{\frac{n+m}{2}} \tag{1}$$

$$CompositionalSim(T_1, T_2) = cosineSim(\sum_{i=1}^{n} wordVector(T_{i1}), \sum_{j=1}^{m} wordVector(T_{j2})) \tag{2}$$

Having retrieved the best GDPR article for the current candidate sentence, the most related paragraph in the identified article should be found (second

step). Lacking a large domain specific corpus, we have modified word2vec model to be able to generate a vector for a sentence and paragraph. According to the literature, a simple yet an efficient way to represent a sentence or a paragraph as a vector is computing the average of all word vectors [16]. In the preparation phase, all `GDPR` paragraph vectors are calculated and stored, whereas the candidate sentence vector is computed in real time. Employing *Cosine* similarity between the candidate sentence vector and all paragraph vectors of the retrieved article, the paragraph with the highest similarity will be identified as the best match.

It is worth to mention that *KnIGHT* is policy and legislation independent. As an example it can be applied to cookie policy which is sometimes embedded into the privacy policy itself. Cookies should comply with the ePrivacy directive[15] that will be soon replaced by proposed ePrivacy regulation[16]. In this case, a corpus of cookies policy in their natural text should be collected to be ingested to *TermRaider*.

4 Evaluation & Discussion

Semantic text mapping is a non-trivial task and its evaluation is just as complex. The ideal assessment method would be to create a gold standard with the help of domain experts; legal experts in this case. For a number of reasons, pursuing this method was not feasible. It was not possible to procure legal experts to perform extremely lengthy tasks (legal policies are lengthy and dense in terms of terminology and implications). Pro bono or voluntary participation from a sufficient amount of experts was also not an option. In addition, legal terms are still rather subjective (and it appears to be markedly more difficult to resolve differences in ideas between legal professionals), and therefore achieving a satisfactory Inter Annotator Agreement (IAA) to generate a gold standard based on which to run the experiment was not possible. For the above reasons, after the first expert concluded (dedicating a total of almost 3 hours) that manual annotation of 4 policies is a time-consuming and subjective task, we changed our strategy and decided to go for a posteriori assessment as our primary experiment. The objective here was to obtain an expert-rated F-measure of the results produced by *KnIGHT*, for the same 4 policies. That said, the primary targeted end-users of the tool are non-experts. Therefore to contextualize these results we conducted the second experiment: 2 lay users were also asked to do what *KnIGHT* does. In this case, to have a realistic yardstick they were instructed to spend between at least 1 and at most 2 hours to go through all 4 policies and identify `GDPR` sections which helped them better understand the makeup of each policy. This exercise, being directly comparable to the first expert's manual annotation, shows the expected success rate of non-expert vs expert `GDPR` mapping.

[15] `http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1525876803065&uri=CELEX:32002L0058`

[16] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:0010:FIN`

4.1 Posteriori Assessment

In order to perform a posteriori assessment, 4 privacy policies from European Union companies were selected and the approach was applied to these policies in their natural language text. Posteriori assessment means running *KnIGHT* over privacy policy texts, finding the matches in `GDPR` and then validating them by legal experts. A successful execution of our pipeline generates some links between privacy policy sentences and `GDPR` paragraphs. Afterwards, in order to reach a semi-conclusive result, 4 legal experts (lawyers or senior law students) were asked to go through the detected links and categorize them into three classes: **related; partially related; or unrelated**[17]. Although *KnIGHT* can generate TOP-n matches of `GDPR` paragraphs for a single candidate sentence, only the best match was considered in the current assessment to reduce the examination time required by experts. In total, 77 annotations were sent to each assessor. Table 1 shows the results per each expert and privacy policy. The *Avg* column denotes the average number for 4 experts per each class, e.g., for *Booking.com* privacy policy, on average 6 out of 23 detected matches were assessed as **related**. However, since privacy and data protection regulations are general in nature and subject to interpretation, there is always a part of subjectivity in legal text assessment. This means that the average column may not necessarily refer to the same annotations for all assessors, e.g., for *Booking.com*, we can not claim that the 6 annotations for **related** class in *Avg* column is the same annotations for all observers.

Table 1. Posteriori Assessment by 4 Experts (E1-E4) for Four Privacy Policies

Privacy Policy	**#Matches**	**Related**					**Partially Related**					**Unrelated**				
		E1	E2	E3	E4	**Avg**	E1	E2	E3	E4	**Avg**	E1	E2	E3	E4	**Avg**
Booking.com	23	7	6	5	6	**6**	12	8	10	9	**9.75**	4	9	8	8	**7.25**
ResearchGate	29	11	14	8	14	**11.75**	8	9	12	4	**8.25**	10	6	9	11	**9**
Ryanair	10	2	3	2	3	**2.5**	4	4	4	3	**3.75**	4	3	4	4	**3.75**
Unilever	15	7	7	2	6	**5.5**	4	4	8	4	**5**	4	4	5	5	**4.5**

IAA is an agreement measure which can be calculated in Kappa or F-meaure. When the observers have the choice to determine the span of the text for annotation, F-measure is recommended [14]. On the other hand, Kappa is appropriate when observers have the same number of classes but with different labels and ranges between -1 and 1 (1:complete disagreement, 0:random agreement, 1:full agreement). Kappa and observed agreements are conventionally computed for two annotators [13]. The extension to more than two annotators is usually taken as the mean of the pair-wise agreements [10]. Furthermore, if the categories (A, B, C, ...) are ordered, weighted Kappa is considered [6]. Our three classes can be treated as an ordered list, because if one expert classifies a match into group

[17] Although we strived to increase the number of evaluators, it was notably hard to find legal experts that agreed to participate in this voluntary task.

related and the other into group **partially related**, this is closer than if one classifies into **related** and the other into **unrelated**.

Tables 2 and 3 show observed agreement and weighted kappa with linear weights. E1 to E4 represent experts and the scores are calculated for all four privacy policies. The results prove that even with a strict number of classes, there is still a part of subjectivity in the assessment and reconfirms the complexity of legal texts. We have provided some examples of agreement and disagreement in table 4. The first sentence from *Booking.com* informs the user that their personal data will only be used with their consent. *KnIGHT* maps this sentence to one of `GDPR` articles about "conditions for consent" and specifically to the paragraph related to the conditions for withdrawing a consent by the data subject. Two experts assessed this match as **partially related**, one as **related** and the other as **unrelated**. Those who annotated this mapping as a partial or perfect match believe that although the sentence is not about withdrawing a consent, the detected `GDPR` paragraph helps the end-user to be aware of their rights. Apart from subjectivity issue, we have realized that the experts tend to have less agreement for short sentences because a short sentence does not say much and it is more controversial. Increasing the window size and considering neighbor sentences will solve this problem. Another issue identified, was generation of incomplete set of tags for some `GDPR` articles. The second sentence in table 4 is mapped to article 77 about "right to lodge a complaint with a supervisory authority" and was labeled as **unrelated** by all experts. This article is a short one with two paragraphs and the generated set of tags contains only three terms: {*supervisory authority, personal data, complaint*}. Therefore the best article retrieval phase detects this article as the best match. This problem can be resolved by narrowing down the domain of the approach, since *KnIGHT* currently uses a general approach without any human involvement. Choosing a specific legislation makes it possible to get help from the domain experts, e.g., in our case we can ask legal experts to manually create some tags for each `GDPR` article. Finally, our evaluations proved that when the similarity score between the candidate sentence and detected paragraph is high, the degree of agreement increases. As an example, the third sentence in table 4 is the best match detected by *KnIGHT* with the similarity equals to 0.75 (max = 1) and it shows complete agreement.

Table 2. Pair-wise Agreement between Experts

Experts	Booking.com	ResearchGate	Ryanair	Unilever
E1 & E2	47.8	79.3	80	66.7
E1 & E3	56.5	62.1	40	53.3
E1 & E4	47.8	82.8	70	73.3
E2 & E3	47.8	62.1	40	60
E2 & E4	52.2	75.9	60	40
E3 & E4	47.8	72.4	50	53.3
Average	**50**	**72.4**	**56.7**	**57.8**

Table 3. Pair-wise Weighted Kappa between Experts

Experts	Booking.com	ResearchGate	Ryanair	Unilever
E1 & E2	30.9	73	76.2	63.4
E1 & E3	38.1	56.5	25	40.6
E1 & E4	34.6	82	63.4	63.4
E2 & E3	33.7	56.8	28.6	43.2
E2 & E4	37.1	66.8	43.2	25
E3 & E4	26.1	70.9	39	43.2
Average	**33.4**	**67.7**	**46**	**46.4**

Table 4. Example of Detected Links & Experts (E1-E4) Assessments (R: related, P: partially related, U:unrelated)

Policy Sentence	Detected `GDPR` Paragraph	E1	E2	E3	E4
Any additional personal details that you give us as a part of the market research will be used only with your consent.	`Article 7(3)`: *The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.*	R	P	U	P
We will comply with all applicable data protection laws and regulations and we will co-operate with data protection authorities.	`Article 77(1)`: *Without prejudice to any other administrative or judicial remedy, every data subject shall have the right to lodge a complaint with a supervisory authority, in particular in the Member State of his or her habitual residence, place of work or place of the alleged infringement if the data subject considers that the processing of personal data relating to him or her infringes this Regulation.*	U	U	U	U
Where a Unilever Site is intended for use by a younger audience, we will obtain consent from a parent or guardian before we collect personal information.	`Article 8(1)`: *Where point (a) of* `Article 6(1)` *applies, in relation to the offer of information society services directly to a child, the processing of the personal data of a child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child. Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.*	R	R	R	R

4.2 Potential End-Users Impact

According to the literature, end-users tend to skip privacy policies and time plays a serious barrier in this case [8]. In order to estimate the time and effort required by end-users for privacy policy comprehension, we asked two non-experts to find the obvious links between four privacy policies of subsection 4.1 and `GDPR`. Here we have used the first expert (E1) annotations (in total 204 links) as a loose gold standard. Table 5 shows the comparison of non-experts annotations and *KnIGHT*'s mapping against E1 gold standard. Since in some cases, the non-experts mapped a single excerpt of a policy to multiple articles, we computed an OR conjunction, e.g., if one of the articles was correct according to E1 gold standard, it was considered as a true positive. As expected, precision and recall are low compared to E1 gold standard and this is inevitable, because experts have a high understanding of privacy policies and in some cases the created links do not have any similar vocabularies but represent an expert inference. On the other hand, the results proves that *KnIGHT* can be a valuable tool for non-experts. Lay end-users spend a lot of time and effort but achieved almost the same F-measure, as opposed to zero effort and instant results of *KnIGHT*.

Table 5. Average F-measure & Total Time of 2 Regular End-users Annotations for 4 Privacy Policies

	Precision	**Recall**	**F1**	**Time (min)**
User1	0.2	0.11	0.14	120
User2	0.46	0.08	0.14	30
KnIGHT	0.3	0.1	0.15	3

4.3 Discussion

The F-scores obtained in subsection 4.2 indicates that there is value in the extraction and mapping method behind *KnIGHT*. On average, based on the experts' ratings between 70-90% of the tool's automatic mappings are at least partially correct (observed agreement with consideration of two classes: partial or perfect match; incorrect match). Of course, the posteriori assessment has its limitations, most notably the lack of consideration for false negatives (missing links). Nevertheless, the results are encouraging more so when considering they are generated instantly whereas typical end-users who performed the annotation task manually - when restricted to 2 hours- only demonstrated an agreement with the expert of just 14% .

Based on the above results, we can conclude that although *KnIGHT* is incomparable to an experts' review of a privacy policy, it does facilitate the mapping of text to relevant articles. As such, it can also be used as a shortcut for both kinds of users alike. For non-experts, it offers a new opportunity for wider awareness

of their rights. Furthermore, it should be stressed out that the number of selected privacy policies and participants in the experiment was a bare minimum. However, we believe that our experimental settings were sufficient to return positive indicative results, ahead of a broader experiment that is in consideration pending sufficient funding.

5 Conclusion & Future Work

In this article, we presented *KnIGHT*, a tool for automatic mapping of privacy policies to GDPR. To the best of our knowledge, this is the first comprehensive approach for privacy policies interpretation and helps regular end-users to familiarize themselves with respective data protection law. Our evaluation showed that interpretation of legal text is a challenging task due to its subjectivity. A comparison of *KnIGHT*'s automatic mapping with two lay end-user annotations proved that *KnIGHT* is able to produce a satisfactory result within a short response time. We deem this work to be a significant step forward to make the regular end-users aware of the their rights as a data subject.

Regarding future work, we aim to make *KnIGHT* a domain specific tool and tailor its workflow for better interpretation of GDPR. Due to the increasing interest in GDPR, there have been initial efforts to represent its article in the form of vocabularies and ontologies [4, 18]. Exploiting the domain specific ontologies enables us to perform a deeper analysis of the texts and extract more knowledge from respective regulations. Furthermore, benefiting from available labeled data set (OPP-115), we seek to improve the candidate sentence detection phase and our matching algorithm.

References

1. Deeplearning4j: Open-source distributed deep learning for the JVM Apache Software Foundation License 2.0 (2015), http://deeplearning4j.org/word2vec.html
2. Acquisti, A., Grossklags, J.: Privacy and rationality in individual decision making. IEEE Security and Privacy **3**(1), 26–33 (Jan 2005). https://doi.org/10.1109/MSP.2005.22, http://dx.doi.org/10.1109/MSP.2005.22
3. Alzahrani, S., Salim, N., Abraham, A.: Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**, 133–149 (2012)
4. Bartolini, C., Muthuri, R.: Reconciling data protection rights and obligations: An ontology of the forthcoming eu regulation. In: Proceedings of the Workshop on Language and Semantic Technology for Legal Domain (LST4LD) (2015)
5. Breaux, T.D., Vail, M.W., Anton, A.I.: Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. In: Proceedings of the 14th IEEE International Requirements Engineering Conference. pp. 46–55. RE '06, IEEE Computer Society, Washington, DC, USA (2006). https://doi.org/10.1109/RE.2006.68, https://doi.org/10.1109/RE.2006.68
6. Cohen, J.M.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological bulletin **70 4**, 213–20 (1968)

7. Costante, E., Sun, Y., Petković, M., den Hartog, J.: A machine learning solution to assess privacy policy completeness: (short paper). In: Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society. pp. 91–96. WPES '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2381966.2381979, http://doi.acm.org/10.1145/2381966.2381979
8. Cranor, L.F., Guduru, P., Arjula, M.: User interfaces for privacy agents. ACM Trans. Comput.-Hum. Interact. **13**(2), 135–178 (Jun 2006). https://doi.org/10.1145/1165734.1165735, http://doi.acm.org/10.1145/1165734.1165735
9. Cunningham, H., Maynard, D., Tablan, V.: JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield (November 2000), http://www.dcs.shef.ac.uk/~diana/Papers/jape.ps
10. Fleiss, J.L.: Measuring agreement between two judges on the presence or absence of a trait. Biometrics **31**(3), 651–659 (1975), http://www.jstor.org/stable/2529549
11. Guntamukkala, N., Dara, R., Grewal, G.W.: A machine-learning based approach for measuring the completeness of online privacy policies. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) pp. 289–294 (2015)
12. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. CoRR **abs/1802.02561** (2018)
13. Hripcsak, G., Heitjan, D.: Measuring agreement in medical informatics reliability studies. Journal of Biomedical Informatics **35**(2), 99–110 (11 2002). https://doi.org/10.1016/S1532-0464(02)00500-2
14. Hripcsak, G., Rothschild, A.S.: Technical brief: Agreement, the f-measure, and reliability in information retrieval. JAMIA (3), 296–298
15. Jensen, C., Potts, C., Jensen, C.: Privacy practices of internet users: Self-reports versus observed behavior. Int. J. Hum.-Comput. Stud. **63**(1-2), 203–227 (Jul 2005). https://doi.org/10.1016/j.ijhcs.2005.04.019, http://dx.doi.org/10.1016/j.ijhcs.2005.04.019
16. Kenter, T., Borisov, A., de Rijke, M.: Siamese CBOW: optimizing word embeddings for sentence representations. CoRR **abs/1606.04640** (2016), http://arxiv.org/abs/1606.04640
17. Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society pp. 1–20 (2018)
18. Pandit, H.J., Lewis, D., O'Sullivan, D.: Gdprtext - gdpr as a linked data resource (Jan 2018). https://doi.org/10.5281/zenodo.1146351, https://doi.org/10.5281/zenodo.1146351
19. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., Norton, T.B., Hovy, E.H., Reidenberg, J.R., Sadeh, N.M.: The creation and analysis of a website privacy policy corpus. In: ACL (2016)
20. Zeni, N., Kiyavitskaya, N., Mich, L., Cordy, J.R., Mylopoulos, J.: Gaiust: Supporting the extraction of rights and obligations for regulatory compliance. Requir. Eng. **20**(1), 1–22 (Mar 2015). https://doi.org/10.1007/s00766-013-0181-8, http://dx.doi.org/10.1007/s00766-013-0181-8