Vol-2385
urn:nbn:de:0074-2385-8

ASAIL 2019
Automated Semantic Analysis of Information in Legal Text

Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts
co-located with the 17th International Conference on Artificial Intelligence and Law (ICAIL 2019)

Montreal, QC, Canada, June 21, 2019.

Edited by

Kevin D. Ashley, University of Pittsburgh, USA
Katie Atkinson, University of Liverpool, UK
L. Karl Branting, MITRE Corporation, USA
Enrico Francesconi, Italian National Research Council (ITTIG-CNR), Publications Office of the European Union
Matthias Grabmair, Carnegie Mellon University, USA
Bernhard Waltl, BMW Group AG, Germany
Vern R. Walker, Maurice A. Deane School of Law at Hofstra University, USA
Adam Zachary Wyner, Swansea University, UK

Table of Contents

Session 6: Machine Learning Method Challenges on Legal Data

2019-06-19: submitted by Matthias Grabmair, metadata incl. bibliographic data published under Creative Commons CC0
2019-06-19: published on CEUR-WS.org |valid HTML5|

GDPR Privacy Policies in CLAUDETTE: Challenges of Omission, Context and Multilingualism

Rūta Liepiņa
Law Department, EUI, Florence, Italy

Giuseppe Contissa
CIRSFID, University of Bologna, Italy

Kasper Drazewski
Law Department, EUI, Florence, Italy

Francesca Lagioia
EUI, Florence, Italy
CIRSFID, University of Bologna, Italy

Marco Lippi
DISMI, University of Modena and Reggio Emilia, Italy

Hans-Wolfgang Micklitz
Law Department, EUI, Florence, Italy

Przemysław Pałka
Yale Law School
New Haven, United States

Giovanni Sartor
EUI, Florence, Italy
CIRSFID, University of Bologna, Italy

Paolo Torroni
DISI, University of Bologna, Italy

Abstract: The latest developments in natural language processing and machine learning have created new opportunities in legal text analysis. In particular, we look at the texts of online privacy policies after the implementation of the European General Data Protection Regulation (GDPR). We analyse 32 privacy policies to design a methodology for automated detection and assessment of compliance of these documents. Preliminary results confirm the pressing issues with current privacy policies and the beneficial use of this approach in empowering consumers in making more informed decisions. However, we also encountered several serious issues in the process. This paper introduces the challenges through concrete examples of context dependence, omission of information, and multilingualism.

1 INTRODUCTION

The changes in online privacy policies following the European General Data Protection Regulation (GDPR) have further highlighted the increasing information asymmetry between online service providers and consumers. Studies [3, 5] in consumer behaviour in reading privacy policies show that long and complex legal documents are seldom read and understood by users. Moreover, [13] show that comprehending the rights and obligations outlined in these online documents is costly both in terms of time and monetary value.

This paper presents a work in progress that includes the latest developments of our methodology [12] in designing the *Gold Standard* of privacy policy compliance that could be used to build a platform empowering consumers to gain easier access and support in understanding their rights and obligations. We aim to provide such a solution through the use of legal analysis, natural language processing, and machine learning. In Section 4, we describe three challenges faced by the AI and Law researchers working on automating evaluation of legal documents and illustrate them through examples found in the privacy policies analysed in our study. Among other issues, we focus on the problem of context dependence of (legal) terms, the challenges in formalising the privacy policies due to their linguistic and legal complexity, and the need for methodologies that can be transferred between different European languages.

In: Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), June 21, 2019, Montreal, QC, Canada.

Published at https://ceur-ws.org

2 BACKGROUND

Legal texts, such as regulations, contracts, privacy policies, and cases, provide a rich source for different formal analyses, due to the complexity of language and legal norms within those texts. One of the aims of artificial intelligence and law research [8, 10] is to find methods for accurately and efficiently extracting the knowledge from legal texts and for providing a level of evaluation for the extracted data. This paper focuses on the legal texts of online privacy policies. We identified three main dimensions for evaluation based on the GDPR and its guidelines: completeness, compliance with the data processing rules, and level of readability. A selection of the research studies in these fields is introduced below.

Completeness: one of the core criticisms against unfair privacy policies regards withheld or missing information on the data processing, such as the purpose and retention time of personal data, including sensitive data. Constante et al. [7] use machine learning and pre-annotated privacy policies to check for the completeness of information pre-GDPR. To this end, they designed a client-end solution, allowing consumers to read summarised policies on privacy categories of their choice (6 core categories and 11 additional categories).

Compliance: service providers, consumers and law enforcement authorities are interested in assessing the compliance of online privacy policies. However, it has proven to be a challenging task. Research in this area focuses on formalising legal norms [4] and designing methodologies [17] for automating the assessment of privacy policies. One of the risks identified [10] relates to the misinterpretation of norms as well as to the failure in connecting different specifications of norms within a legal document.

Readability: a different area of research focuses on the language and accessibility of privacy policies. A new study [5] provides empirical evidence on the readability levels of privacy policies post-GDPR, concluding that "these policies are often unreadable".[1] Following previous work by [14], their results support the conclusion

[1] For readability scores the study employed the Flesch Reading Ease (FRE) test and the Flesch-Kincaid (F-K) test.

that an unreasonable level of expertise is required to comprehend the privacy policies. The average score, among the 300 analysed policies, was at a level of "the usual score of articles in academic journals" [5], supporting the claim that policies are not written to be accessible and understandable by the general public. Such barriers further discourage consumers from reading privacy policies [16]. Some solutions, such as automatically generated privacy policy summaries [19] and interactive solutions of privacy analysis through apps [1], are emerging to provide consumers with tools to better understand the contents of agreements and exercise their rights.

3 DESIGNING METHODOLOGY

This project aims to design a methodology for creating an open and high quality annotated corpus of online privacy policies. Such a data set could be used for automated detection and evaluation of problematic privacy clauses given the GDPR as the basis for integrated normative guidelines. Here, we present an overview of the current methodology for detecting and assessing the problematic privacy clauses, and how the new guidelines have improved on previous versions [6].

3.1 The Gold Standard

We designed a methodology that reflects the overall aims of the GDPR in regards to collection and processing of personal data. In particular, we focus on three ways a privacy policy can be deemed unlawful according to articles 13 and 14 of the GDPR: (1) if the policy omits information required by the regulation, (2) if the policy defines data processing beyond the prescribed limits, and (3) if it is written in unclear language.

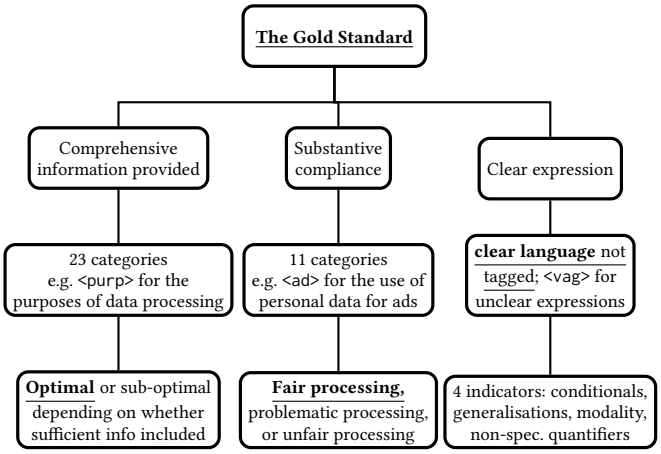

Figure 1: Dimensions - categories - criteria

We chose three top-level dimensions for the evaluation:

(1) comprehensiveness of information
(2) substantive compliance
(3) clarity of expression

Each of the top-level dimensions has been further divided into the relevant categories and corresponding evaluation criteria. Diagram 1 shows the layered structure of the methodology by exemplifying a good privacy policy: one that satisfies all the criteria.[2] To meet the requirements of comprehensiveness, a privacy policy should declare the purposes of the processing precisely and exhaustively. Thus, clauses providing only examples must be considered as insufficiently informative. In the dimension of substantive compliance, using personal data for targeted advertising is fair only if based on the data subject's consent and whenever an opt-out is possible. Regarding the clarity of expression, i.e. whether a privacy policy is framed in understandable, precise, and intelligible language, certain unspecific language qualifiers should be avoided (e.g. indeterminate conditioners, creating a dependency of a stated action or activity on a variable trigger such as "as necessary", "from time to time", etc). We have designed detailed annotation guidelines that are being further tested with a new data set of policies.

(1) Comprehensiveness of Information. The clause satisfies the criteria if the privacy policy includes sufficient information on the 23 categories defined in the annotation guidelines. These include: `<id>` identity of the data controller, `<cat>` categories of personal data concerned, and `<ret>` the period for which the personal data will be stored. Where 'sufficiency' is defined as fully informative privacy clauses that include all the details required by the regulation (e.g. `<id1>`). Everything that does not satisfy the given criteria, as specified in the guidelines has been marked as sub-optimal (e.g. `<id2>`). We use the numerical values of 1 and 2 in the XML tags to refer to the level of comprehensiveness of the information given. The earlier version of the methodology distinguished 12 relevant categories. The number of categories was increased to 23 to provide a more fine-grained annotation of functions. The improvements from the previous annotation guidelines [6] consist of the further specification of the different functions of the rights granted to consumers, and the steps needed to exercise them. In particular, the clauses implementing the duty to inform the data subject about their rights, under article 13.2(b) and 14.2(c) of the GDPR, initially falling under a single category of required information[6] and identified with the `<correct>` tag, have been distinguished in multiple categories. The reason for further differentiating between such categories is twofold. Firstly, from the legal point of view, the right to request access to, and rectification or erasure of, personal data or restriction of processing and to object to processing, as well as the right to data portability, are conceptually distinct and independent. Secondly, in analysing the privacy policies, we noted that the different rights and steps needed to exercise these rights are usually addressed in separate clauses. Thus we chose the units for our tagging method as single phrases. Indeed, with clauses covering multiple sentences, we chose to tag each sentence separately, by treating statements independently from one another. Hence, also the clauses containing information about the rights are now classified separately from those outlining the steps needed to exercise these rights. Consider, for instance, the following example:

> `You can request access to your personal information, or correct or update out-of-date`

[2] In the diagram, the **underlined** criteria illustrate a good privacy policy.

```
or inaccurate personal information we hold
about you. You can most easily do this by
visiting the "Account" portion of our web-
site, where you have the ability to access
and update a broad range of information
about your account, including your contact
information, your Netflix payment informa-
tion, and various related information about
your account (such as the content you have
viewed and rated, and your reviews.
```

Under the previous version of the tagging guidelines, the two clauses, considered separately, were not deemed as exhaustive with regard to the initial <correct> category and were marked as *insufficiently informative* (for instance, the first clause fails to inform the data subject about the existence of the right to object to processing, as well as about the right to data portability). In the example below, we illustrate how we now further distinguish <acc> for the right to request access to personal data from the data controller, <corr> the right to request the rectification of personal data, <cat> the categories of personal data concerned, and <sacc> the steps needed to exercise the right to access their personal data.

```
[Current version]<acc2><corr2><cat2>You can
request access to your personal information,
or correct or update out-of-date or inac-
curate personal information we hold about
you.</cat2>
</corr2></acc2>
<sacc1><acc1><corr1>You can most easily do
this by visiting the "Account" portion of
our website, where you have the ability to
access and update a broad range of infor-
mation about your account, including your
contact information, your Netflix payment
information, and various related informa-
tion about your account (such as the con-
tent you have viewed and rated, and your
reviews).</corr1></acc1></sacc1>
```

The 23 category guidelines for comprehensiveness of information are currently being tested against the hypothesis that the added categories will enhance the precision of answers given to the consumers.

(2) Substantive Compliance. In dimension of substantive compliance, we distinguish 11 categories of clauses pertaining to the types of processing. A clause is considered fair if the defined data processing practices are permitted by, and thus compliant with, the GDPR (Art.5, 6, and 9). We assumed that each clause can be classified either as a *fair processing* clause <tag1>, *problematic processing* <tag2>, or *unfair processing* <tag3> clause. We used the numerical values of 1, 2, and 3 for each XML tag to indicate the level of fairness. In this dimension, the two levels of sub-optimal achievement of the Gold Standard distinguish between problematic clauses, where it may be reasonably doubted that the clause meets the GDPR requirements, and unfair clauses, where the data processing clearly fails to meet the GDPR requirements, i.e. the data processing defined in the policy document is forbidden by the regulation.

We identified 11 categories of clauses based on how issues pertaining to such categories might affect individual rights. For instance, the unfair processing of sensitive (<sens>) data, or unauthorised transfer of data to third parties (tp) can have negative consequences for the consumer. Other categories pertain to the consent by using practice, the take it or leave it approach, policy changes and whether there has been a fair warning, cross-border data transfer, consent for processing children's data, licensing data, advertising, any other types of consent, as well as one category for tracking any other types of problematic clauses.

(3) Clarity of Expression. Art 12 specifies that a privacy policy should be framed "in a concise, transparent, intelligible and easily accessible form, using clear and plain language". To integrate this requirement into the assessment criteria, four indicators for vagueness (categories of linguistic expressions possibly generating indeterminacy, depending on the context) were defined [18]: (1) indeterminate conditioners, creating a dependency of a stated action or activity on a variable trigger, such as "as necessary", "from time to time", etc.; (2) expression generalisations, abstracting actions and activities under unclear conditions and contexts, such as "generally", "normally", " largely", "often", etc.; (3) modality, including adverbs and non-specific adjectives, which create uncertainty with respect to the possibility of certain actions and events, and (4) nonspecific numeric quantifiers, creating ambiguity as to the actual measure of a certain action and activity, such as "numerous", "some", "most", "many", "including (but not limited to)", etc. Note that a single clause may fall into different categories, in different dimensions, and consequently may have multiple tags. For example, if the clause allows for a problematic processing of sensitive data and includes vague terms, it is marked as:

```
<sens><vag>The sentence.</vag></sens>
```

3.2 A Preliminary Corpus

In the privacy policy assessment, we worked with a corpus of 32 policies, manually tagged by two independent annotators. Privacy policies were selected on the basis of the number of users and the platform's global relevance, as well as taking into account our previous work [6, 12] analysing Terms of Services for the same online services. We used XML mark-up language for annotations.

The data set contains 6,275 sentences. As we observed above, the sentences were tagged according to 35 categories (23 under the comprehensiveness of information dimension, 11 under substantive compliance, and 1 under clarity of expression). In the remainder of the paper we will only mention some of these categories and we will report on experiments concerning three categories (<purp>, <ad>, and <vag>): one for each dimension of the Gold Standard defined in Section 3.1. <purp> for the comprehensiveness of information, <ad> for substantive compliance, and finally <vag> for unclear language. The corpus contains 773 sentences tagged with <purp>, out of which 281 and 492 sentences refer to cases of sufficient (<purp1>) and partial (<purp2>) information, respectively. As for advertising, 91 sentences in the corpus are tagged as problematic (<ad2>) whereas 95 are tagged as unfair (<ad3>). Finally, 714 sentences are tagged as unclear (<vag>).

We hereby remark that, in this paper, we are presenting a preliminary version of the corpus for which the tagging guidelines

directed to annotators have been revised multiple times. We plan to make these guidelines stable and publicly available in the near future, once the corpus is finalised. At that stage, we also intend to measure the inter-annotator agreement in order to assess the quality of the deployed data set.

4 CHALLENGES

In this section, we describe the challenges that we envision when aiming to develop an automatic system for the assessment of compliance of privacy policies according to the GDPR. All examples have been extracted from the Airbnb Privacy Policy document, last updated 16 April 2018.

4.1 Context

One of the earliest challenges encountered in the automated detection of problematic clauses in privacy policies is the fact that the examination of single sentences is insufficient for the determination of their defectiveness within the three dimensions. For this purpose we need to link several sentences. Conversely, our previous experiments showed that the analysis of single sentences is adequate to identify unlawful or unfair clause in terms of services. For instance, consider the following example taken from the Airbnb privacy policy.

> [Line 80] 2.2 Create and Maintain a Trusted and Safer Environment. Detect and prevent fraud, spam, abuse, security incidents, and other harmful activity.
> Conduct security investigations and risk assessments.
> Verify or authenticate information or identifications provided by you (such as to verify your Accommodation address or compare your identification photo to another photo you provide).
> Conduct checks against databases and other information sources, including background or police checks, to the extent permitted by applicable laws and with your consent where required.
> Comply with our legal obligations.
> Resolve any disputes with any of our Members and enforce our agreements with third parties.
> Enforce our Terms of Service and other policies.
> In connection with the activities above, we may conduct profiling based on your interactions with the Airbnb Platform, your profile information and other content you submit to the Airbnb Platform, and information obtained from third parties. In limited cases, automated processes may restrict or suspend access to the Airbnb Platform if such processes detect a Member or activity that we think poses a safety or other risk to the Airbnb Platform, other Members, or third parties.
> We process this information given our legitimate interest in protecting the Airbnb Platform, to measure the adequate performance of our contract with you, and to comply with applicable laws.

As it can be seen, the last sentence taken separately fails to specify the legitimate interest at stake, the specification there provided "protecting the Airbnb Platform, to measure the adequate performance of our contract with you, and to comply with applicable laws", which is very generic. However, the sentence offers an adequate specification when it is read in conjunction with the preceding list. This means that for the detector to identify defectiveness of a clause, it should evaluate the whole section, rather than the individual sentences.

4.2 Omission of Information

In our previous work [12] on Terms of Service, we used machine learning and natural language processing techniques for the detection of (potentially) unfair clauses. In the context of privacy policies we have different goals, which are defined in the Gold Standard guidelines (see Section 3.1). In particular, our purpose lies not only in detecting the unfairness, and the unclear language,[3] but also in checking whether certain information is present and sufficient in view of the regulatory framework.

The latter is conceptually a completely different task for two main reasons: (i) we aim to identify the *presence* of a sentence, rather than the fact that its content is not compliant with the law, and (ii) we need to verify whether some information is *sufficient*, or not, with respect to the Gold Standard.

In case of Terms of Service, classic NLP approaches, such as statistical classifiers or neural networks, worked quite well since the detection of unfair clauses can be easily framed as a sentence classification problem, where (potential) unfairness is clearly defined and statistics collected from a wide corpus can be sufficient to identify target clauses. In contrast, in the privacy policy analysis our goal is not pure detection of content, since it also involves the capability to spot some missing, hidden, or insufficient information. For humans, this problem is typically addressed with a number of reasoning steps. Therefore, we argue that more sophisticated artificial intelligence approaches are needed, for example coming from the neural-symbolic community [9], or from the neural architectures that have been specifically developed to deal with reasoning tasks [11]. Another path for development could be explored by adding contextual information to the classifier. For instance, when classifying a single sentence, taking into account also the information regarding surrounding sentences, or even the whole document, could in fact provide crucial information for a correct classification of the clause.

As an example of the complexity of such a task, we hereby report some clauses related to the purpose of processing (<purp>) within the comprehensiveness dimension. Following the GDPR, the data controller is required to provide clear information on the purposes

[3]The detection of unclear language is also *per se* a slightly different task, as it moves the attention towards a purely linguistic perspective.

as to *why* data are collected and *how* such data will be used. These processes should be transparent and within the limits prescribed in articles 13(1)(c) and 14(1)(c). To assess whether the privacy policy is compliant in this regard, we distinguish between optimal (fully informative) and sub-optimal (missing some information) clauses.

For example the following clause satisfies the criteria since it provides an exhaustive list of the purposes for data processing.

> <purp1>If you are a Host, the Payments Data Controller may require identity verification information (such as images of your government issued ID, passport, national ID card, or driving license) or other authentication
> information, your date of birth, your address, email address, phone number and other information in order to verify your identity, provide the Payment Services to you, and to comply with applicable law.</purp1>

In contrast, clauses that use vague language and only give general examples are considered problematic, since they can be interpreted to justify the use of personal data beyond what the consumer might have intended when consenting to the policy. It raises concerns around informed consent. Consider, for instance the following example from the Airbnb Privacy Policy.

> <purp2>We may use your personal data to develop new services</purp2>

4.3 Multilingualism

Considering that the GDPR governs data processing in all European Union states, it is important to take into account its 24 official languages. Linguistic diversity and equal legal status between the different European languages are among the core values in access to justice in the EU. Therefore, when offering any solution aimed at informing and protecting consumers, researchers should also design its methodology to preserve the original functions and accuracy across these many different languages. This task is particularly relevant for NGOs and consumer organisations that very often struggle with the diversity of language and the comparison of different versions of the same documents.

In our project, we have chosen English as the base language, and have started experimenting with transfer of tags from annotated documents in English to privacy policies in German. This process involves the use of three types of documents: (1) the original, annotated text in English, (2) the original text in German, and (3) the automatic translation of the original English text into German.

Consider, for instance, the following examples of original, annotated clauses in English. The first clause pertains to the period for which the personal data will be stored. It has been marked as <ret2>, i.e. insufficiently informative, since it does not clearly define the retention period of the personal data. The second clause pertains to both the data retention and the categories of data collected. It has been marked as insufficient since the retention period and the categories of personal data are not defined, as indicated by the expressions 'reasonable measures' and 'when it is no longer required'.

> [ENGLISH] <ret2>We may retain information as required or permitted by applicable laws and regulations, including to honor your choices, for our billing or records purposes and to fulfill the purposes described in this Privacy Statement.</ret2>
> <ret2><cat2>We take reasonable measures to destroy or de-identify personal information in a secure manner when it is no longer required.</cat2></ret2>

Let us now consider the corresponding clauses in German as translated and marked.

> [GERMAN] <ret2>Wir können Informationen, wie gemäß geltenden Gesetzen und Bestimmungen erforderlich oder zugelassen, einschließlich unter Einbeziehung ihrer Auswahl, zu zwecken der Rechnungstellung oder Buchführung und um den zwecken dieser Datenschutz Erklärung nachzukommen, speichern.</ret2>
> <cat2><ret2>Wir ergreifen angemessene Maßnahmen, um personenbezogene Daten auf eine sichere Weise zu zerstören oder unkenntlich zu machen, wenn diese nicht länger erforderlich sind.</ret2></cat2>

In this test case, the machine translation reference file was generated in an accurate manner and the tags were successfully transferred, given that the English and German language versions did not bear discrepancies in the clauses used.

Clearly, there would be major challenges involved with transferring tags in cases where the text in English is different from the text in target language, not only in terms of syntax, but also regarding the legal obligations that might be unique to a certain jurisdiction. Moreover, English is by far the most widely studied language in natural language processing, thus the existing resources in other languages are often not as accurate or rich as those developed for English. Nevertheless, a lot of effort in artificial intelligence is currently being dedicated to tools and platforms dealing with multilingualism (e.g., see [2, 15] and references therein).

5 EXPERIMENTS

In this section we present some preliminary results, based on the data set of 32 annotated privacy policies, as described in Section 3.2. We focus on the task of sentence detection only, leaving to future work the challenges related to multilingualism.

In particular, in our experimental evaluation we used SVMHMM, a machine learning approach that combines Support Vector Machines (SVM) and Hidden Markov Models (HMM) [20], and which enables to collectively classify all the sentences in a document, thus taking into account the order of the examples. We started with a very basic set of features, namely the bag-of-words (unigrams and bigrams) describing each sentence, leaving to future research a deeper investigation of richer feature sets, possibly exploiting deep learning in order to directly learn sentence representations.

In all the experiments we used the leave-one-document-out (LOO) procedure, where each document is used, in turn, as the

Table 1: Macro-averaged results achieved by SVMHMM on the LOO setting. To highlight the difficulty of the task, we also report the performance of a random predictor, and a trivial classifier always predicting the positive class.

Tag	Method	P	R	F_1
<ad>	SVMHMM	0.408	0.565	0.421
	Random	0.034	0.034	0.034
	Always Positive	0.032	1.000	0.061
<purp>	SVMHMM	0.602	0.586	0.552
	Random	0.126	0.126	0.126
	Always Positive	0.126	1.000	0.221
<vag>	SVMHMM	0.412	0.612	0.460
	Random	0.112	0.112	0.112
	Always Positive	0.112	1.000	0.196

test set, and all the remaining are merged into the training set. We consider the following performance measures: (i) precision P, that is the fraction of sentences predicted as positive, which are actually positive; (ii) recall R, that is the fraction of positive sentences that are correctly detected; (iii) F-measure F_1, that is the harmonic mean between P and R. For each measure, we report the macro-average, that is the average computed over the measures obtained for each single document.

We consider the tasks of detecting the clauses concerning the purpose of processing (thus considering the union of <purp1> and <purp2> as the positive class), those problematic or unfair related to advertising (with the union of <ad2> and <ad3> as the positive class), and finally those that contain unclear language (the <vag> tag only). Results are reported in Table 1. To highlight the difficulty of the task, we compare the results achieved by SVMHMM against two trivial baselines: a random classifier, which predicts the positive class accordingly to class distribution, and a second system that always predicts the positive class. SVMHMM achieves a value of F_1 equal to 0.552 for the detection of clauses regarding the purpose of processing (against 0.126 and 0.221 of the two baselines, respectively) and 0.421 for advertising (against 0.034 and 0.061 for the two baselines, respectively). A similar trend is shown for unclear language, which achieves F_1 equal to 0.460. The very low values of the baselines, as well as the confusion matrices reported in Table 2, clearly show the large imbalance between the positive and negative classes: for example, only 3% of sentences are annotated as either <ad2> or <ad3>. This imbalance makes all the considered tasks particularly challenging. Therefore the F_1 values obtained in the range $0.42 - 0.55$ can be considered as encouraging.

In addition, we also want to note that the results are very heterogeneous across different documents. For example, for the <ad> tag, for the Dropbox and Courchsurfing policies, the SVMHMM approach achieves F_1 equal to 0.86 and 0.89, respectively, whereas the Crowtangle policy is even perfectly predicted, with three positive clauses correctly predicted with no false positive. We plan to deeply analyse and discuss further these more fine-grained results once our final corpus will be released.

6 DISCUSSION AND FUTURE WORK

Considering the number of independent research projects working in this area, an identification of the current problems aims to establish a common ground for fruitful discussions of the future work. In this paper, we have presented a work in progress of a methodology (the Gold Standard) for annotating post-GDPR privacy policies to identify and assess the compliance with the regulation. We have identified three challenges that should be addressed to progress in assessing the privacy policies with NLP and ML tools. While we have made some progress in each of the identified areas, there remains a lot of work to reach the overall objectives of the project.

The first challenge concerns the fact that the privacy policies are written in a language that tends to be more broad in its possible interpretations, and it is not uncommon to define the meaning of certain terms early in the document and use such terms without direct references back to the original definitions. Such references can be both internal and external, increasing the complexity for comprehension of the consumer's rights and duties based on the signed agreement. Since our project aims at providing consumers with a tool that would facilitate an increased understanding of the privacy policies, it is essential that the automated evaluation of clauses is able to build context for such an understanding.

The second challenge focused on the omission of information, which requires both the knowledge of what information should be included in the document and a way to identify the absence of the required information. Such a task requires exploring methods beyond pure text mining approaches.

Lastly, we looked at the need to consider an approach that is able to use the results achieved in working with privacy policies in English and transfer the annotations to different language versions without losing the accuracy and efficiency.

In sum, with ever more scientific research going open-access, the need for clear and transparent annotation guidelines and shared corpora is increasingly pressing. As part of our future work, we aim to publish the annotated privacy policy corpora online, as we have done with the Terms of Service agreements. Future work also includes moving beyond pure language processing and introducing a level of reasoning that allows context comprehension by machines. We maintain our overall objective to design a methodology and provide a tool for consumers and NGOs that would empower them through more informed decision making in the digital environment.

7 ACKNOWLEDGEMENTS

We would like to thank all the members of the Project Claudette and our funding authorities at the European University Institute Research Council, Bureau Européen des Unions de Consommateurs, and the Zeppelin Universität.

REFERENCES

[1] Lisa M Austin, David Lie, Peter Yi Ping Sun, Robin Spillette, Michelle Wong, and Mariana D'Angelo. Towards dynamic transparency: The apptrans (transparency for android applications) project. *http://dx.doi.org/10.2139/ssrn.3203601*, 2018.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. Does anyone read the fine print? consumer attention to standard-form contracts. *The Journal of Legal Studies*, 43(1):1–35, 2014.

		Predicted	
		0	1
True	0	5,893	196
	1	88	98

		Predicted	
		0	1
True	0	5,199	303
	1	327	446

		Predicted	
		0	1
True	0	4,969	592
	1	297	417

Table 2: Micro-averaged confusion matrices for the three considered detection tasks: <ad> (left), <purp> (center), and <vag> (right). The positive class (1) represents sentences of that specific tag, whereas the negative class (0) represents all the other sentences. Large class imbalance is evident in all cases. Note that the precision and recall metrics obtained from these tables slightly differ from the results in Table 1 because here we are reporting micro-average rather than macro-average.

[4] Cesare Bartolini, Gabriele Lenzini, and Cristiana Santos. A legal validation of a formal representation of gdpr articles. *In CEUR Workshop Proceedings:*, http://ceur-ws.org/Vol-2309/10.pdf.

[5] Shmuel I Becher and Uri Benoliel. Law in books and law in action: The readability of privacy policies and the gdpr. *CONSUMER LAW & ECONOMICS, Klaus Mathis & Avishalom Tor, eds., Springer (forthcoming, 2019)*, 2019.

[6] Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans-W Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. *https://ssrn.com/abstract=3208596*, 2018.

[7] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. A machine learning solution to assess privacy policy completeness:(short paper). In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, pages 91–96. ACM, 2012.

[8] Mauro Dragoni, Serena Villata, Williams Rizzi, and Guido Governatori. Combining nlp approaches for rule extraction from legal documents. In *1st Workshop on MIning and REasoning with Legal texts (MIREL 2016)*, 2016.

[9] Artur d'Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. Neural-symbolic learning and reasoning: contributions and challenges. In *2015 AAAI Spring Symposium Series*, 2015.

[10] Mustafa Hashmi. A methodology for extracting legal norms from regulatory documents. In *2015 IEEE 19th International Enterprise Distributed Object Computing Workshop*, pages 41–50. IEEE, 2015.

[11] Herbert Jaeger. Artificial intelligence: Deep neural reasoning. *Nature*, 538(7626):467, 2016.

[12] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, pages 1–23, 2018.

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

[14] George R Milne, Mary J Culnan, and Henry Greene. A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing*, 25(2):238–249, 2006.

[15] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[16] Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, pages 1–20, 2018.

[17] Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. Pronto: Privacy ontology for legal reasoning. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 139–152. Springer, 2018.

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

[19] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 163–166. International World Wide Web Conferences Steering Committee, 2018.

[20] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM, 2004.