

# **Knowledge Extraction Methods for the Analysis of Contractual Agreements**

Dissertation  
zur  
Erlangung des Doktorgrades (Dr. rer. nat.)  
der  
Mathematisch-Naturwissenschaftlichen Fakultät  
der  
Rheinischen Friedrich-Wilhelms-Universität Bonn

von  
**Najmehsadat Mousavinezhad**  
aus  
Babolsar, Iran

Bonn, 19.10.2020

Dieser Forschungsbericht wurde als Dissertation von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem publication server of the University of Bonn <https://nbn-resolving.org/urn:nbn:de:hbz:5-64537> elektronisch publiziert.

1. Gutachter: Prof. Dr. Sören Auer  
2. Gutachter: Prof. Dr. Jens Lehmann

Tag der Promotion: 31.05.2021  
Erscheinungsjahr: 2021

# Abstract

---

The ubiquitous availability of the Internet results in a massive number of apps, software, and online services with accompanying contractual agreements in the form of ‘end-user license agreement’ and ‘privacy policy’. Often the textual documents describing rights, policies, and conditions comprise many pages and can not be reasonably assumed to be read and understood by humans. Although everyone is exposed to such consent forms, the majority tend to ignore them due to their length and complexity. However, the cost of ignoring terms and conditions is not always negligible, and occasionally people have to pay (money or other means) as a result of their oversight.

In this thesis, we focus on the interpretation of contractual agreements for the benefit of end-users. Contractual agreements encompass both the privacy policies and the general terms and conditions related to software and services. The main characteristics of such agreements are their use of legal terminologies and limited vocabulary. This feature has pros and cons. On one hand, the clear structure and legal language facilitate the mapping between the human-readable agreements and machine-processable concepts. On the other hand, the legal terminologies make the contractual agreement complex, subjective, and, therefore, open to interpretation. This thesis addresses the problem of contractual agreement analysis from both perspectives.

In order to provide a structured presentation of contractual agreements, we apply text mining and semantic technologies to develop approaches that extract important information from the agreements and retrieve helpful links and resources for better comprehension. Our approaches are based on ontology-based information extraction, machine learning, and semantic similarity and aim to deliver tedious consent forms in a user friendly and visualized format. The ontology-based information extraction approach processes the human-readable license agreement guided by a domain ontology to extract deontic modalities and presents a summarized output to the end-user. In the extraction phase, we focus on three key rights and conditions: *permission*, *prohibition*, *duty*, and cluster the extracted excerpts according to their similarities. The clustering is based on semantic similarity employing a distributional semantics approach on large word embeddings database. The machine learning method employs deep neural networks to classify a privacy policy’s paragraphs into pre-defined categories. Since the prediction results of the trained model are promising, we further use the predicted classes to assign five risk colors (Green, Yellow, Red) to five privacy icons (*Expected Use*, *Expected Collection*, *Precise Location*, *Data Retention* and *Children Privacy*). Furthermore, given that any contractual agreement must comply with the relevant legislation, we utilize text semantic similarity to map an agreement’s content to regulatory documents. The semantic similarity-based approach finds candidate sentences in an agreement that are potentially related to specific articles in the regulation. Then, for each candidate sentence, the relevant article and provision is found according to their semantic similarity. The achieved results from our proposed approaches allow us to conclude that although semi-automatic approaches lead to information loss, they save time and effort by producing instant results and facilitate the end-users understanding of legal texts.



# Acknowledgements

---

Throughout the exciting Ph.D. journey, I met several people who inspired and supported me all those years. I would like to thank Prof. Dr. Sören Auer for giving me a chance to pursue a degree in Germany at the Enterprise Information Systems department at the University of Bonn. His supervision and advice fostered my whole research. He also supported me a lot for getting involved in industry projects. Furthermore, I would like to thank Prof. Dr. Jens Lehmann for his invaluable help and constant support regarding the bureaucratic procedures of DAAD scholarship. Moreover, I would like to thank my advisor Dr. Simon Scerri, without whom I could not win the DAAD scholarship! He introduced me to everyday scientific routines, which inspired me to do academic research and provided me with many fruitful scientific discussions. I would also like to thank Prof. Dr. Maria-Esther Vidal whose valuable comments made me a better researcher.

Last but not least, I thank my family and friends for their unconditional support and would like to express my gratitude to my colleagues, Dr. Steffen Lohmann, Dr. Christoph Lange, Dr. Giulio Napolitano, Dr. Ioanna Lytra, Dr. Diego Collarana, Dr. Sahar Vahdati, Dr. Michael Galkin, Dr. Damien Graux, Rostislav Nedelchev, Debanjan Chaudhuri, Mohamed Nadjib Mami, Isaiah Onando Mulang', Fathoni A. Musyaffa, Elisa Margareth Sibarani, Afshin Sadeghi, Mehdi Ali, Gëzim Sejdiu, Mohnish Dubey, Mirette Elias and Sebastian Bader for motivating me by their cheerfulness and passion and encouraging me to achieve better results.



# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Problem Statement and Challenges . . . . .	2
1.3	Research Questions . . . . .	4
1.4	Thesis Overview . . . . .	5
1.4.1	Contributions . . . . .	5
1.4.2	Publications . . . . .	7
1.5	Thesis Structure . . . . .	9
<b>2</b>	<b>Background</b>	<b>11</b>
2.1	Different Types of Click-Wrap Agreements . . . . .	11
2.1.1	End-User License Agreement . . . . .	11
2.1.2	Privacy Policy . . . . .	12
2.1.3	Terms of Use . . . . .	13
2.2	Semantic Representation and Ontologies . . . . .	13
2.2.1	Resource Description Framework . . . . .	13
2.2.2	Ontologies . . . . .	15
2.3	Machine-Processable Contractual Agreement Representation . . . . .	15
2.3.1	Rights Expression Language . . . . .	15
2.3.2	Policy Languages . . . . .	18
2.4	Deep Learning Foundations . . . . .	20
2.4.1	Multilayer Neural Network . . . . .	20
2.4.2	Convolutional Neural Network . . . . .	22
2.5	Background Overview . . . . .	24
<b>3</b>	<b>Related Work</b>	<b>25</b>
3.1	Legal Texts Enrichment using Semantic Annotations . . . . .	25
3.1.1	Manual Enrichment . . . . .	25
3.1.2	Phrase Heuristics (Linguistic Rules) . . . . .	31
3.2	Interpretation of Contractual Agreements using Machine Learning . . . . .	33
3.2.1	Linear Classification Methods . . . . .	33
3.2.2	Deep Neural Networks . . . . .	35
3.3	Toward Violation Detection in Enterprise Policies . . . . .	36
3.3.1	Semantic Similarity Based Approaches . . . . .	36
3.3.2	Machine Learning Based Approaches . . . . .	37

<b>4</b>	<b>Semantic Interpretation of Contractual Agreements using Ontologies</b>	<b>39</b>
4.1	Ontology-Based Information Extraction from License Agreements	40
4.1.1	Architecture	40
4.1.2	GATE OBIE Pipeline	41
4.1.3	Word Space Creation and Semantic Clustering	45
4.1.4	<i>EULAide</i> Framework and Web Service	46
4.2	Experimental Study	47
4.2.1	OBIE Pipeline Evaluation	48
4.2.2	Evaluating the Clustering Approach	52
4.2.3	Usability Experiments	55
4.3	Summary	57
<b>5</b>	<b>Analysis of Contractual Agreements using Deep Learning</b>	<b>59</b>
5.1	Background: OPP-115 Dataset	60
5.2	Establishing a Baseline for Privacy Policy Classification	63
5.2.1	Pre-trained Word Embeddings	63
5.2.2	Convolutional Neural Network	64
5.2.3	Bidirectional Encoder Representations from Transformers	65
5.2.4	<i>Pripolis</i> Framework	66
5.3	Risk Level Prediction	66
5.4	Experimental Study	67
5.4.1	Multi-label Classification Evaluation	67
5.4.2	Risk Icons Evaluation	71
5.5	Summary	72
<b>6</b>	<b>Mapping Contractual Agreements to Regulatory Documents</b>	<b>75</b>
6.1	Mapping Privacy Policies to the GDPR	77
6.1.1	Pipeline Preparation	78
6.1.2	Semantic Text Matching	79
6.2	Experimental Study	81
6.2.1	Posteriori Assessment	81
6.2.2	Potential End-Users Impact	84
6.3	Summary	84
<b>7</b>	<b>Conclusion</b>	<b>87</b>
7.1	Research Questions Analysis	87
7.2	Limitations	90
7.3	Closing Remarks and Future Work	90
	<b>Bibliography</b>	<b>93</b>
	<b>A Low-level Attributes of <i>Pripolis</i></b>	<b>105</b>
	<b>List of Figures</b>	<b>109</b>
	<b>List of Tables</b>	<b>111</b>



---

## Introduction

---

In this Chapter, we will introduce the work done in the thesis, first by motivating the research problem, then stating the challenges and finally giving an overview of the thesis structure.

### 1.1 Motivation

The ubiquitous nature of the Internet resulted in an ever-growing number of online and mobile services for end-users, ranging from personal information management (e.g., Webmail, calendar, address book), cloud storage (e.g., photo/video repositories) over collaboration tools (e.g., document authoring, messaging) to e-commerce (online shops, song/movie subscription services). Every day new services emerge, and their providers aim at quickly increasing the user base and market share by providing user-friendly interfaces; frequently even permitting users to use the service completely free of charge. In most cases, users have to accept terms and conditions governing the usage before utilizing such services. However, their use remains regulated through specific terms and conditions, and not infrequently, users are unaware of their obligation to ‘pay’ for the service by sharing their personal data and contributions. Figure 1.1 shows excerpts from *ResearchGate*’s consent forms<sup>1</sup>. The terms of service includes the general conditions for using the website as well as the copyright policy concerning the materials posted on *ResearchGate*. In addition, the privacy policy explains how *ResearchGate* collects and uses the user’s data in compliance with the General Data Protection Regulation (GDPR) [1].

The problem arises when people ignore the consent forms due to their length and complicated terminology. In the 2016 study, “*The biggest lie on the Internet*”, 543 students were asked to agree to the privacy policy and terms of use in order to join a fictitious social network [2]. Although 26% did not choose the ‘quick join’ button, the average time of reading was only 73 seconds. Ignoring these terms is a risk taken by most users. According to another survey conducted by *Skandia* [3], 10% of people said that they were bound by a longer contract than they expected, and 5% lost money by not being able to cancel or amend their bookings.

In order to facilitate the process of digesting consent forms for regular end-users, we consider applying text mining and semantic technologies to provide important information in a structured scheme. Semantic technologies are envisioned for extracting knowledge from the raw data sources and forming semantic networks. They assist machines to understand data. As one of the building blocks of semantic technology, ontologies are used to describe domain concepts and relationships between them in a machine-readable language. Since contractual agreements are a type of legal contract, they exhibit the same characteristics

---

<sup>1</sup>To retrieve the exact source used: <<https://www.researchgate.net/terms-of-service>>, <<https://www.researchgate.net/privacy-policy>> – last accessed September.2<sup>nd</sup>.2020

Terms of Service	Privacy Policy
<ol style="list-style-type: none"> <li>1. General information</li> <li>2. Conclusion of agreement</li> <li>3. Scope of the Service</li> <li>4. User Obligations</li> <li>5. Export control laws</li> <li>6. Copyright and intellectual property rights</li> <li>7. Unsolicited ideas</li> <li>8. Changes to the Service</li> <li>9. Changes to these Terms</li> <li>[...]</li> </ol>	<ol style="list-style-type: none"> <li>1. Introduction</li> <li>2. Information we process and how we process it</li> <li>3. How we use data we collect about you</li> <li>4. How we use relevant publicly available data</li> <li>5. Information we process when using technologies like cookies and pixels</li> <li>6. Advertisements on ResearchGate</li> <li>7. Third party services for analytics, measurement and ad delivery</li> <li>[...]</li> </ol>
(a) The terms and conditions for using the service.	(b) The privacy notice.

Figure 1.1: Excerpts from *ResearchGate* contractual agreements.

of legal agreements, e.g., they are written in a clear structure and use legal terminologies. This attribute facilitates the ‘mapping’ of natural-language text to machine-readable conceptualizations in the ontology for our use-cases.

In addition to semantic technologies, we exploit deep learning, which has shown huge success in natural language processing. Deep learning is a subset of machine learning which uses several layers in the neural network. The network requires a labeled dataset to learn the “hidden” features of the input data. If the dataset contains enough samples and the provided labels are reliable, the trained model can predict promising results for the ‘unseen’ input data. In the legal domain, it is of paramount importance that the input data is annotated by domain experts. Hence, we exploited a highly endorsed dataset that was created and annotated by a group of experts.

The approaches considered in this thesis are broadly applicable to other forms of text-based contractual agreements. By changing training data/vocabulary and tailoring the domain-specific rules, other kinds of agreements can potentially be consumed by our architecture. However, in this thesis, we specifically address End-User License Agreements (or EULAs) and privacy policies for proof of concepts, since they have the broadest impact and affect everyone.

## 1.2 Problem Statement and Challenges

In the digital age, everyone is exposed to terms and conditions regulating the use of services and software, and in their majority, this constitutes ordinary people with limited to no knowledge of legal terms. The major problem arises when the end-user ignores the contractual agreements due to their length and difficulty. Therefore, this thesis contributes to the practicable approaches for *interactive presentation of contractual agreements using knowledge extraction methods*.

In order to interpret and analyze the consent forms for the benefit of regular end-users, we face multiple challenges. First, due to the complex terminology of such texts, one must rely on the domain knowledge which is produced by experts. Therefore, we explore the usability of domain vocabularies and ontologies as well as reliable datasets, created by the experts. In *Challenge 1*, knowledge extraction methods are investigated to study their suitability for extracting valuable information from contractual agreements. Second, given that our approach is designed to be consumed by the regular end-users, we investigate

whether the extracted information assists users to spend less time digesting consents forms (*Challenge 2*). Finally, considering the fact that all contractual agreements should comply with the applicable laws, we explore the feasibility of mapping them to the relevant legislation (*Challenge 3*).

As the main problem is much larger than it can be seen from the above descriptions and can not be easily solved with one thesis, we leave out of the thesis scope numerous tasks and challenges, e.g., compliance checking of human-readable agreements with the help of legal experts, investigating other kinds of legal agreements, providing more domain-specific resources and particular applications. We are convinced that the new findings presented in this thesis would serve a promising basis for future work addressing those out of the scope challenges.

### **Challenge 1: Extracting Valuable Information from Contractual Agreements**

The attributes of contractual agreements require special investigation. First, as a type of legal contract, they tend to have complex terminology. In an empirical study conducted by two law professors, 500 of the most popular websites in the United States were analyzed, and the sign-in wrap contracts that these sites use were studied [4]. On Average, the readability score was comparable to the usual score of articles found in academic journals (14.1 years of education). Of the 500 websites “Terms of Service” agreements, more than 100 required a reading level that is even higher than 14.1 years of education. The complex content of agreements restricts the range of people who are able to interpret such texts properly. Hence, in this work, we must rely on vocabularies and datasets that are created by domain experts, which in some cases, are very rare to find and exploit.

In addition to the complexity, most consent forms are at least a few pages long. According to a survey, if users were to read the privacy policies of all services they visit on the Internet, they would need to spend on average 244 hours each year, which is almost over half of the average time a user spends on the Internet [5]. This characteristic makes the evaluation of our approach very challenging. Most people are reluctant to participate in those experiments where they have to interact with a very long and tedious text. Considering the contractual agreements’ features, one of our main challenges is extracting information that, from the regular user’s perspective, is useful and beneficial.

### **Challenge 2: Efficient Presentation of Information from the End-User Perspective**

As the primary target users of our thesis are the regular end-users (as opposed to legal experts or lawyers), one of our main concerns is presenting the extracted information efficiently. Usability is a crucial concept in any user-centered design. According to [6], usability is the “Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.” Our designed services must be simple and convenient for immediate use by regular end-users, without the need for ‘How-to’s’ and instructions. The goal of this thesis is motivating people to make themselves familiar with the conditions they are agreeing to as well as the type of personal data they are sharing with service providers. Therefore, another challenge for us is designing a service that is easy to use and helps the users to digest contractual agreements.

### **Challenge 3: Subjective Interpretability of Legal Texts**

Legal interpretation, like interpretation in general, fluctuates between objectivity and total subjectivity. Objectivity aims to eliminate decisions based on personal bias, where, according to Balkin, “Subjectivity is what the individual subject brings to the act of understanding; it is what allows her to construct the object of her interpretation so that she can understand it” [7].

On one hand, as the legal movement emphasizes, interpretation is always at least partly subjective [8]. On the other hand, the responsible entities attempt to minimize the subjective elements when defining norms and regulations. However, contractual agreements are designed by internal experts in enterprises. Small companies may not even afford to hire lawyers to create such policies. Hence, the published agreements are frequently very subjective in interpretation and in some cases unintentionally infringe the applicable laws. As a result, the interpretation of contractual agreements relies heavily on the values of the person making the interpretation and, therefore, is a constant challenge in legal text processing.

### 1.3 Research Questions

To solve the above challenges in this thesis we resort to the use and application of domain ontologies, semantic similarity and deep learning. Based on the challenges, we derive the following research questions:

#### Research Question 1 (RQ1)

Are text mining techniques able to extract valuable information from contractual agreements?

Natural language processing (NLP) at the high-level is categorized into rule-based approaches and statistical methods. Rule-based techniques rely on hand-coded rules defined and written by humans, whereas statistical approaches apply machine learning algorithms. In this thesis, we investigate both approaches. For the rule-based information extraction, we rely on a domain-specific ontology and design our rules based on the ontology's specifications. Here, our hypothesis is that since contractual agreements have a clear structure and terminologies, definition of rules is more straightforward (in comparison to arbitrary texts) and that the 'mapping' between human-readable agreements and machine-processable concepts in the ontology is feasible. Moreover, we employ supervised deep learning to explore statistical approaches. Convolutional neural network and transformers are two types of neural networks that have led to outstanding performance in the last few years [9–13]. As a part of this research question, we study the suitability, and performance of deep learning techniques in the legal domain and more precisely, for privacy policy interpretation and analysis.

#### Research Question 2 (RQ2)

Does ontology-based information extraction help end-users to spend less time to understand contractual agreements?

In this question, our initial hypothesis is that the information loss produced by (semi-)automatic approaches is a reasonable cost for the time saved by users. The goal is to estimate our method's usability, from the end-user perspective. Thus, we design several experiments to verify that the extracted information helps end-users digest the contractual agreements and that our approach is a proper replacement for the long and sophisticated natural language policies. We strive to provide user-friendly services that are easy to comprehend and beneficial for the end-users. The designed services intend to encourage people to study consent forms before accepting them.

### Research Question 3 (RQ3)

Given the subjective interpretability of legal texts, to what extent can we map contractual agreements to the applicable laws?

Policies and agreements are meaningless without the law, and all of them must comply with the relevant regulations. Considering the subjectivity of legal text interpretation, we study the possibility of mapping contractual agreements to the applicable law, to help the end-users familiarize themselves with their rights as a licensee (the entity who accepts the license). For this research question, we identify some unavoidable limiting factors. First, there is, and will always be, a part of subjectivity in such an experiment. Most rules are general in nature and subject to interpretation, which does not facilitate their classification into rigid categories. Second, it is hard to decide on what triggers the link between an excerpt from a contractual agreement and a specific provision in the regulation. When it comes, for instance, to the right of access in the GDPR, several provisions may prove relevant: (1) the reasons why the right of access is mentioned in the privacy policy and (2) the modalities surrounding its exercise. Considering these factors, we strive to design an experiment to explore the feasibility of finding relations between an agreement and the regulatory documents.

## 1.4 Thesis Overview

To present a high-level but descriptive overview of the achieved results during the conducted research, we emphasize the main contributions of the thesis and provide references to scientific articles covering these contributions published throughout the whole term.

### 1.4.1 Contributions

#### Contributions for RQ1

Extracting permissions, obligations, and prohibitions from End-User License Agreements (EULAs) and interpretation of privacy policies using deep learning.

We investigate ontology-based information extraction to extract deontic modalities from end-user license agreements. The ODRL ontology is exploited to annotate the license agreements with the ontology concepts. The embedded ontology-aware gazetteer is able to provide the ontological class of each entry and find mentions in the text matching instances in the ontology. The matching can be done between any morphological or typographical variant (e.g., upper/lower case, CamelCase). Having annotated ontological instances, we then define several linguistic rules based on the ontology's specifications. In addition to extracting important excerpts, a hierarchical clustering approach is applied to categorize similar deontic modalities. Although OBIE is a standard approach, there has been no prior study utilizing OBIE for EULAs. From this point of view, our application is new. Furthermore, since we are benefiting from a standard 'model' of the domain, there is a huge potential to better structure similar documents along with the same taxonomy. Moreover, having a vocabulary to cover such legal texts can become a standard for structuring also new documents (and not just the existing ones).

To study statistical methods for natural language processing, neural network is employed to classify privacy policy paragraphs into pre-defined categories (which were specified by legal experts). First, the state-of-the-art results are reproduced and then using a powerful framework, the results are further improved. Furthermore, we use the predicted categories by the deep learning module to assign five risk colors to five privacy icons. The conducted experiments against a reliable gold standard show that our results are promising and beneficial. In the absence of a standard baseline in this area, our contribution for privacy policy classification can be considered as a strong candidate.

#### Contributions for RQ2

Implementation of *EULAide* and *Pripolis* to facilitate the end-user interaction.

Since the primary target audience of this thesis are the regular end-users, we design and implement user-friendly interfaces to assist the users in understanding consent forms. More specifically, in the matter of EULAs, a set of qualitative and quantitative experiments are conducted. In order to evaluate the efficiency of *EULAide* we conducted an experiment to identify if the solution enables end-users to invest less time to sufficiently comprehend license agreements. At the same time, we wanted to identify the trade-off between the added support and the information loss expected when applying semi-automatic IE and stigmatization. The experiments were designed to identify how well regular people can remember policies and how fast they can search for information in an EULA. In practice, when one is agreeing with terms, this process should be followed so as to be aware of the rights and regulations. The results verify our initial hypotheses, i.e., even though *EULAide* is effected by information loss, it considerably saves time and effort spent by users to arrive to a similar level of understanding. Although the number of selected EULAs and participants was the bare minimum required for an experiment of this kind (due to funding restrictions), the results were still sufficient to indicate value in extending and improving our approach.

#### Contributions for RQ3

Implementation of *KnIGHT* (Know your rIGHT) for mapping privacy policies to the General Data Protection Regulation (GDPR).

In order to take the initial step for compliance checking of an agreement's text, we present a general approach based on the 'text semantic similarity' to relate license agreements to the applicable laws in sentence and paragraph level. As a proof-of-concept, we apply our approach to privacy policies. Since privacy policies are stipulating how companies will gather, manage, and process customer data, they must comply with the data protection laws. The General Data Protection Regulation (GDPR) applies since 25 May 2018, and specifically has the purpose of making significantly easier for citizens to have control over their personal data. In addition it aims at enforcing organizations to respect the data subject's rights, e.g., data subject can have any data which is stored with one service provider transmitted directly to another provider (data portability). Our approach finds the relations between a privacy policy and the GDPR articles. Such relations assist user to know their rights as a citizen. For instance, the generated links inform users that they have the right to request a copy of their information without any cost, or that they can contact companies and request not to use their data in the marketing activities. Such mappings opens

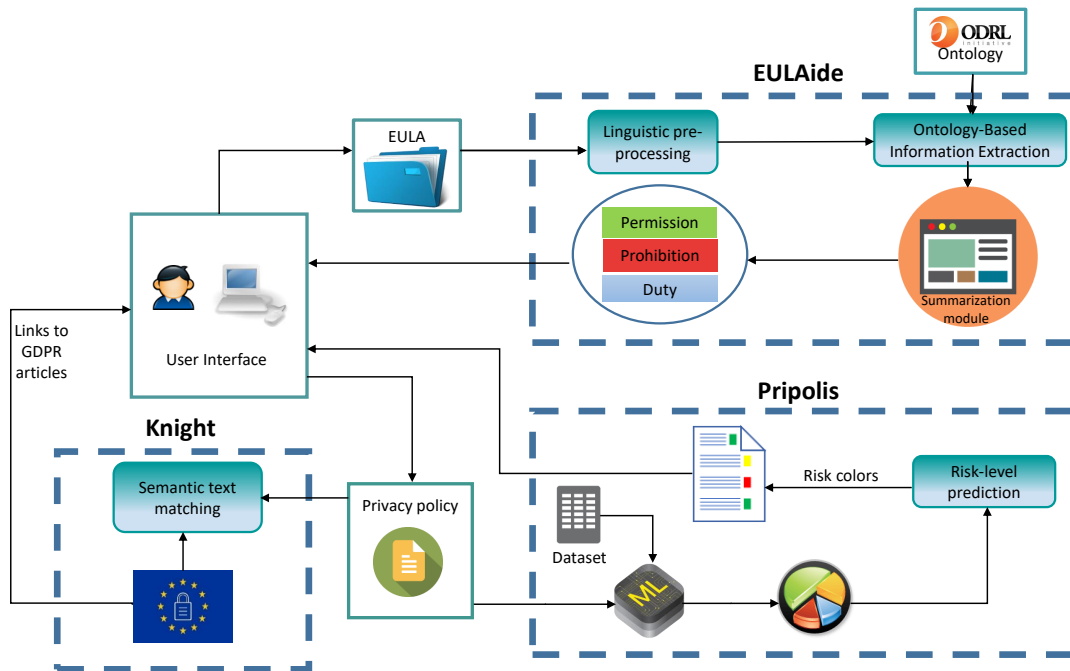


Figure 1.2: The high-level workflow of our approach.

a door towards the automatic compliance checking of agreements. However, since (semi-)automatic methods always lead to information loss, the designed approaches can never replace an expert’s role in detecting potential violation, instead, they can serve as a recommender system.

## 1.4.2 Publications

Figure 1.2 illustrates the systematic representation of how our approach works from a high-level perspective. It consists of three separate modules. *EULAide* is responsible for analyzing license agreements and is founded on OBIE. This module takes the human-readable license agreement and ODRL ontology as inputs and produces a condensed summary of permissions, prohibitions and duties. Similarly, *Polisis* is trained with the OPP-115 dataset and the labeled data. It takes a human-readable privacy policy and predicts pre-defined categories based on the trained model. In addition, it performs a shallow risk analysis based on the predicted categories and assigns five risk colors (green, red, yellow) to five privacy icons (*Expected Use*, *Expected Collection*, *Precise Location*, *Data Retention* and *Children Privacy*). Finally, *KnIGHT* maps a privacy policy’s paragraphs to the GDPR’s articles. The mapping algorithm uses word embeddings and semantic similarity between texts, to find the best GDPR articles to the privacy policy content. *KnIGHT* aims at informing regular end-users about their rights as a data subject by providing useful links and resources.

The following publications constitute a scientific basis of this thesis and serve as a reference point for numerous figures, tables, and ideas presented in the later chapters:

1. **Najmeh Mousavi Nejad**, Simon Scerri, Sören Auer, Elisa Margareth Sibarani. *EULAide: Interpretation of end-user license agreements using ontology-based information extraction*. In Proceedings of the 12th International Conference on Semantic Systems, SEMANTiCS 2016, pages 73–80, ACM; In this paper, I presented an ontology-based information extraction method for EULA



term and phrase extraction to facilitate a better understanding by humans. An ontology capturing important terms and relationships has been used to guide the OBIE process. In the detection and extraction, I focused on three key rights and conditions: `permission`, `prohibition` and `duty`. I name my approach *EULAide*, which comprises a custom information extraction pipeline and a number of custom extraction rules tailored for EULA processing. This paper was *nominated* for the **Best Research and Innovation Paper Award**.

2. **Najmeh Mousavi Nejad**, Simon Scerri, Sören Auer. *Semantic Similarity based Clustering of License Excerpts for Improved End-User Interpretation*. In Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, pages 144-151, ACM; This paper is the continuation of the previous publication. I improved *EULAide* by clustering similar extracted excerpts (`permission`, `prohibition` and `duty`) in order to ease the process of license analysis for end-users. The clustering is based on semantic similarity employing a distributional semantics approach on large word embeddings database. Furthermore, I implemented *EULAide* as a web service that can be communicated by any client.
3. **Najmeh Mousavi Nejad**, Pablo Jabat, Rostislav Nedelchev, Simon Scerri, Damien Graux. *Establishing a Strong Baseline for Privacy Policy Classification*, In Proceedings of 35th International Conference on ICT Systems Security and Privacy Protection, IFIP-SEC 2020, pages 370-383, Springer; In this paper, my contribution is establishing a strong baseline for privacy policy classification. I presented three different models that are able to assign pre-defined categories to privacy policy paragraphs, using supervised machine learning. In order to train the neural networks, a dataset containing 115 privacy policies were exploited. I published the implementation and all resources openly to ensure that my achieved results are easily reproducible.
4. **Najmeh Mousavi Nejad**, Damien Graux, Diego Collarana. *Towards Measuring Risk Factors in Privacy Policies*. In Proceedings of the Workshop on Artificial Intelligence and the Administrative State co-located with 17th International Conference on AI and Law (ICAIL 2019), pages 18-20; Founded on the previously mentioned baseline, in this paper, I proposed to measure a policy's risk factor based on the predicted categories and attributes. For those pre-defined classes that the trained model shows low accuracy and F-measure (due to the scarce number of samples), I proposed to define hand-coded rules using experts' annotations. Given the clear and structured terminology of privacy policies, the rule-based extraction method yields promising results.
5. **Najmeh Mousavi Nejad**, Simon Scerri, Jens Lehmann. *KnIGHT: Mapping Privacy Policies to GDPR*. In Proceedings of 21st International Conference on Knowledge Engineering and Knowledge Management, EKAW 2018, pages 258-272, Springer; In light of the, now enforced EU-wide, General Data Protection Regulation (GDPR), I proposed *KnIGHT* (Know your rIGHTs), 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* 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 paragraphs. The conducted experiments show that with further improvement, it is feasible to design a recommender system that assists legal experts to find potential violations in privacy policies.
6. **Najmeh Mousavi Nejad**. *Semantic Analysis of Contractual Agreements to Support End-User Interpretation*. In Proceedings of the EKAW Doctoral Consortium 2018 co-located with the 21st International Conference on Knowledge Engineering and Knowledge; The EKAW Doctoral Consortium is an opportunity for Ph.D. students to discuss and obtain feedback on their ongoing



work, plans, and research directions with/from experienced researchers in the field. The objective is to share best practices of research methods and approaches, as well as to exchange on what it means to engage in an academic and research career on the topics relevant to the EKAW conference. Students accepted to present at the Doctoral Consortium attended the Doctoral Consortium for the whole day. Among 12 presentations, I won the **best presentation award** of the consortium.

## 1.5 Thesis Structure

The thesis is structured into seven chapters. Chapter 1 introduces the thesis covering the main research problem, motivation for the conducted study, research questions, scientific contributions that address research questions, and a list of published scientific papers that formally describe those contributions.

Chapter 2 presents fundamental concepts and preliminaries that serve as foundations of our research and are necessary for understanding the rationale behind our chosen approaches.

Chapter 3 discusses state-of-the-art community efforts in various domains, e.g., legal text enrichment using semantic annotations, machine learning based approaches for interpretation of contractual agreements, and violation detection in companies' policies.

In Chapter 4, we introduce *EULAide*, a framework that applies ontology-based information extraction to extract deontic modalities from end-user license agreements. In addition to extracting important excerpts from EULAs, *EULAide* clusters the similar extracted segments based on semantic similarity and provide a basic summary for each cluster. A comprehensive set of qualitative and quantitative experiments are conducted to evaluate the performance of *EULAide*.

Chapter 5 reports the efforts carried out to interpret and analyze privacy policies. Despite the presence of a reliable dataset for privacy policies, there is no standard benchmark in literature, and therefore, we introduce a strong baseline for privacy policy classification. The conducted experiments show that our approach successfully reproduces state-of-the-art and further improves the results.

Chapter 6 delves into mapping privacy policies to GDPR. Since all contracts and agreements should comply with the applicable laws, we present a general approach to find the relations between agreements and the relevant regulatory documents. *KnIGHT* (Know your rIGHT) is an effort to assist end-users to familiarize themselves with their rights as a data subject.

Finally, Chapter 7 concludes the thesis with the directions of future work. We once more look through the research questions and provide answers using the obtained results.



# Background

---

In this Chapter, we present basic concepts that serve as foundations of the research conducted in this thesis. In Section 2.1, we first introduce different types of contractual agreements used as our main use-cases in the upcoming sections. As the thesis discusses ontology-based information extraction, in Section 2.2, we then briefly explain semantic technologies. In Section 2.3, we cover the discussion of machine-processable languages, which are designed explicitly for expressing rights, obligations, and policies. Section 2.4 discusses deep learning foundation and finally, in Section 2.5, we provide a summary of the background topics covered in this Chapter.

## 2.1 Different Types of Click-Wrap Agreements

A contractual agreement is a form of contract that restricts access, defines the use, and ensures protection of the involved parties. According to [14], “A click-wrap agreement is a digital prompt that offers individuals the opportunity to accept or decline a digitally-mediated policy”. Privacy policies, Terms of Service (ToS) (also known as Terms of Use (ToU) and Terms and Conditions (T&C)) and copyright policies usually employ the click-wrap prompt, since they often require clicking with a mouse on an icon or a button to accept the agreement. The copyright policies are regularly embedded in the end-user license agreements. In the upcoming subsections we briefly explain each type of click-wrap agreements.

### 2.1.1 End-User License Agreement

A software license agreement is commonly called an End-User License Agreement (or EULA). EULA is a legal contract that governs the use or redistribution of the software. Under copyright law, all software are a type of literary work<sup>1</sup>, and therefore are copyright protected [15]. EULAs must comply with the applicable laws, and the law determines if the rights are acceptable. Some of the rights, protected by the copyright law of the European Union are [16]:

- *right of reproduction*;
- *right of distribution*;
- *right of rental* and/or *lending*.

---

<sup>1</sup>For the definition of literary work for the purpose of copyright law in Germany, see [https://www.gesetze-im-internet.de/englisch\\_urhg/englisch\\_urhg.html](https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html)

In addition to copyright policies, EULAs typically contain clauses that define liability and responsibility between the parties entering into the agreement. The common provisions in EULAs are:

- **License Granting:** grants a license of use for an app to the end-user;
- **Restriction of Use:** prevents any illegal use of the app;
- **Related Agreement:** notifies the user that agreeing to the license may lead to adherence to the terms of other agreements;
- **Copyright Infringement/Intellectual Property:** states that the user will be held responsible for any legal issues in the case of copyright infringement;
- **Termination of Licensing:** grants rights to the provider or licensor of the app to terminate the license in the event of a violation of use or other issues;
- **Warranty Disclaimer:** clarifies that the licensor is not responsible for improving the app to satisfy the end-users;
- **Limitations of Liability:** explains that the licensor will not be held responsible for any damages that may arise from the use of the software.

For further reading, we refer the reader to [17].

### 2.1.2 Privacy Policy

Privacy policies are legal documents stipulating how companies will gather, manage and process customer data. They are legally required for any service that uses, maintains, or discloses data that can be used to identify an individual, e.g., name, date of birth, contact information, address, and many more. In contrast to EULAs, privacy policies must comply with a smaller set of legislation, i.e., data protection and privacy laws. Since May 2018, all privacy policies in the European Union should comply with the GDPR [1]. Website privacy policies are also dependent on “Privacy and Electronic Communications Directive” [18].

Privacy policies typically contain broader and more generalized statements rather than only representing the company’s data usage policy. According to a few studies, the content of privacy policies had a considerable amendment after the GDPR [19–21]. Common provisions, according to the GDPR are:

- information that is collected and how it is collected;
- how the collected information is used (purpose);
- whether the personal information will be shared with third parties;
- ensuring the security of information processing;
- how long the personal data is retained;
- data subject rights;
- contact information of the service provider.

For further reading, we refer the reader to [22].

### 2.1.3 Terms of Use

Terms of use (also known as “terms and conditions” and “terms of service”), is a legal agreement between a service provider and the end-user. Terms of use is not equivalent to EULAs, e.g., they cover a broader content and mainly clarify how the end-user should behave while using the service. The service can be whether an app, software, websites, social networks, search engines, credit cards, file storage, and other types of services. As a result, depending on the type of service, terms of use must comply with different legislation (general laws applicable to Internet technology and content, e-commerce law, commercial law, export control laws, etc.). For instance, if a website’s target end-users are foreign consumers, the requirements of foreign laws and the possibility of being sued in a foreign country must be considered [23]. Common provisions in terms of use are:

- **Definitions/General Information:** definition of keywords and phrases;
- **Using the Service:** the user’s permissions and obligation regarding the use of the service;
- **Jurisdiction and dispute resolution:** terms in the case of arising an international jurisdiction, venue for legal disputes and consumer dispute resolution;
- **Changes to the terms or service:** reserving the right to change the service features and terms of use at any time and how notices will be sent to the consumers;
- **Disclaimers and Limitations of Liability:** states that the service provider is not responsible for enhancing the service to satisfy the consumers’ needs and that the end-users will be liable for any kind of damage that may arise out of the use of the service (to the extent that is permitted by applicable law);
- **Privacy & Cookies:** Privacy information and link to a Privacy Policy

We should note here that due to the broadness of terms of use agreements, they are not the focus of this thesis.

## 2.2 Semantic Representation and Ontologies

Semantic technologies aim to understand the meaning of data by introducing open standards for describing data and information. In this Section, we explain the most important standards that are known as the vital backbones of semantic technologies.

### 2.2.1 Resource Description Framework

Resource Description Framework (RDF) is a data model for data interchange on the Web and is a W3C recommendation since 1998 [24]. Although XML is able to model data and information, there is no unique way to represent knowledge in XML. RDF was introduced to solve this problem.

RDF was originally used to describe metadata for web resources, and then have been generalized to encode structured information. It is based on the form of **subject–predicate–object** (or **entity–attribute–value**), known as triples. The **subject** defines the resource or the asset and can be either a URI or a blank node. Uniform Resource Identifiers or URIs are used to reference resources unambiguously, and a blank node is used to represent an individual with certain properties without a name. The **predicate** or **property** denotes the relationship between the **subject** and the **object** and is always an URI. Finally, the **Object** could be URIs, blank nodes, or literals (data values).

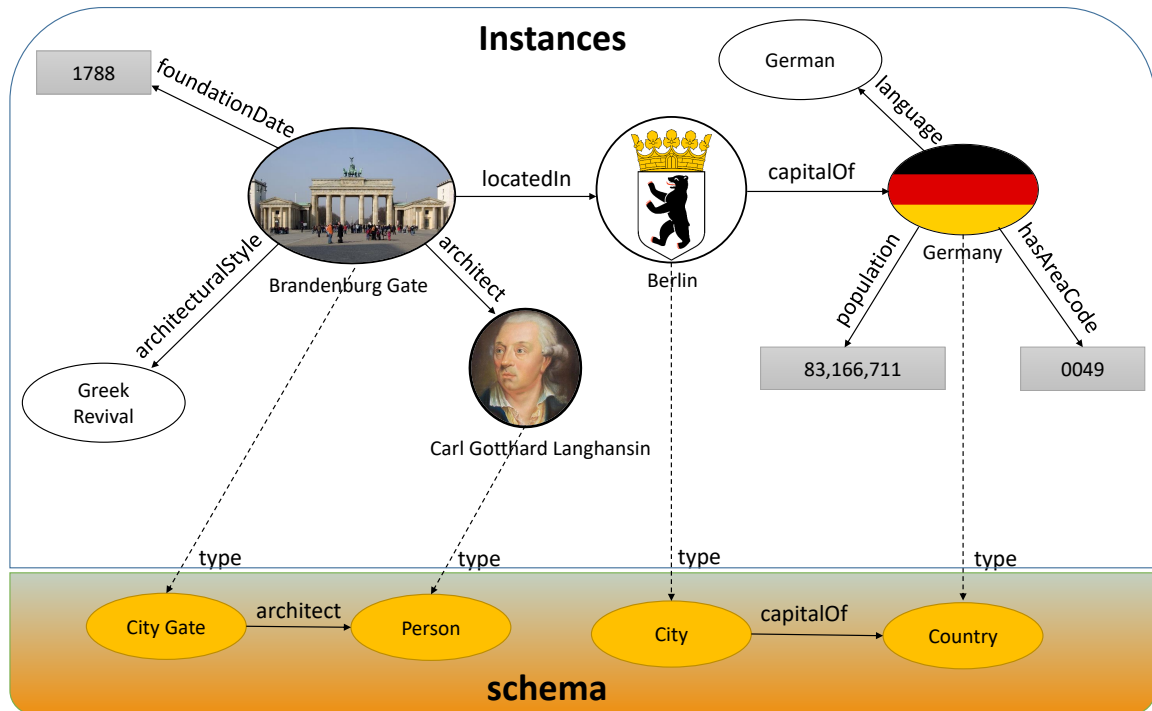


Figure 2.1: An example of RDF graph. Resources are denoted by circles and literals are denoted by rectangles.

Since RDF is not able to represent a schema, one of its limitations is the possibility of defining irrational triples. Therefore, RDF Schema extends RDF with a special Vocabulary for terminological knowledge (as opposed to RDF assertional knowledge). RDFS (S = “Schema”) is a part of the W3C RDF recommendation family and puts constraints on the use of RDF by allowing to define classes and properties [24]. It is an RDF vocabulary; thus, every RDFS graph is an RDF graph. In RDFS, we can define a set of things or entities by **rdfs:Class**. Furthermore, using the properties **rdfs:domain** and **rdfs:range** we can define the domain and range of a property.

Figure 2.1 demonstrates an example of a set RDF triples. The orange area shows the schema, and the instances are presented in the upper area (know as Abox). In this example, *CityGate*, *Person*, *City* and *Country* are RDFS classes and the instance are linked to these classes by **rdf:type** attribute. Assuming all resources of the figure belong to the ontology of namespace ‘ex’ (except ‘type’), the triples of the RDF graph is presented in listing 2.1.

Last but not least, there are different serialization formats for RDF:

- **Turtle**: a text format known for its human readability;
- **N-Triples**: a text format focusing on simple parsing;
- **Notation 3**: or N3 is a text format with advanced features beyond RDF;
- **RDF/XML**: the official XML [25] serialization of RDF;
- **JSON-LD**: the official JSON [26] serialization of RDF;
- **RDFa**: a mechanism for embedding RDF in HTML.

Listing 2.1 uses Notation 3 serialization which is also known for its human readability.

```

1 @prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2 @prefix ex:    <http://www.example.org/> .
3 @prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
4
5 ex:Brandenburg_Gate      rdf:type          ex:CityGate .
6 ex:Carl_Gotthard_Langhans rdf:type          ex:Person .
7 ex:Berlin                rdf:type          ex:City .
8 ex:Germany               rdf:type          ex:Country .
9 ex:Brandenburg_Gate      ex:foundationYear "1788"^^xsd:string .
10 ex:Brandenburg_Gate      ex:architecturalStyle ex:Greek_Revival .
11 ex:Brandenburg_Gate      ex:architect          ex:Carl_Gotthard_Langhans .
12 ex:Brandenburg_Gate      ex:locatedIn          ex:Berlin .
13 ex:Berlin                ex:capitalOf      ex:Germany .
14 ex:Germany               ex:language      ex:German_language .
15 ex:Germany               ex:population  "83166711"^^xsd:integer .
16 ex:Germany               ex:hasAreaCode "49"^^xsd:integer .

```

Listing 2.1: RDF graph of 2.1 represented in RDF (Notation 3 syntax).

## 2.2.2 Ontologies

In the previous part, we saw that RDF Schema can be used to define a "lightweight" vocabulary. However, RDF Schema has some limitations regarding the possibilities of formulating ontologies. For instance, it is not possible to specify that a property's domain should not contain a certain class. Moreover, there is no way to define cardinalities and metadata of the schema.

Ontology languages allow us to capture the meaning of information by specifying how information interacts with other information using their formal semantics. Ontologies consist of a set of axioms that can be expressed as a set of RDF triples. According to Gruber [27], "An Ontology is a formal specification of a shared conceptualization of a domain of interest." The Web Ontology Language or OWL (more easily pronounced than WOL), is a family of languages for authoring ontologies. As opposed to RDF Schema, OWL allows to instantiate classes by individuals, provides means to define concept inheritance and transitivity, symmetry, functionality, and inverse functionality for properties. Furthermore, it contains logical class constructors such as **owl:intersectionOf** for conjunction, **owl:unionOf** for disjunction, and **owl:complementOf** for negation. In the next Section, we study the domain-specific vocabularies and ontologies that are used to express contractual agreements.

## 2.3 Machine-Processable Contractual Agreement Representation

In recent years, there has been a growing interest in generating machine-readable contractual agreements. Consequently, several machine-readable languages have been proposed. In the sequel, we introduce two established languages expressing the rights and policies of agreements.

### 2.3.1 Rights Expression Language

Rights Expression Language (REL) is a machine-readable language that declares rights and permissions. RELs are expressible in different languages, such as XML, RDF, RDF Schema, and JSON. Among

these languages, RDF has drawn much attention over the past years. A REL structure based on **entity-attribute-value** may contain:

- **Entities**: such as *Things, Classes, Work, Asset, License, End-User, Party* or *Jurisdiction*;
- **Attributes**: properties that belong to each entity, e.g., for a *License*, common attributes are: *Permissions, Prohibitions, Duties* and *Constraint*;
- **Values**: values of these properties, e.g., for a *Permission*, some values are: *copy, distribute, display*.

In this section, we focus on the three established expression languages that are widely used in the Semantic Web community.

## CC REL

Creative Commons Rights Expression Language (CC REL) explains how license information can be expressed in a machine-readable format using RDF. For each *cc:License*, CC REL has a set of properties:

- **cc:permits**: an action that may be allowed, e.g., *cc:Reproduction, cc:Distribution, cc:DerivativeWork, cc:Sharing*;
- **cc:prohibits**: an action that the user is not allowed to do, e.g., *cc:CommercialUse*;
- **cc:requires**: an action that the user must fulfill in order to be granted a certain permission, e.g., *cc:Notice, cc:Attribution, cc:ShareAlike, cc:SourceCode, cc:Copyleft, cc:LesserCopyleft*.

Listing 2.2 shows the CC Attribution-NonCommercial license<sup>2</sup> represented in RDF (Notation 3 syntax<sup>3</sup>), where permissions are *Reproduction, Distribution*, and *Derivative Works*, requirements are *Notice* and *Attribution*, and *Commercial Use* is prohibited. For more details on the CC REL vocabulary, we refer the reader to [28].

```

1 @prefix cc:      <http://creativecommons.org/ns#> .
2 @prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3 @prefix ex:     <http://www.example.org/> .
4
5 ex:licCC-BY-NC  rdf:type          cc:License .
6 ex:licCC-BY-NC  cc:permits       cc:Reproduction .
7 ex:licCC-BY-NC  cc:permits       cc:Distribution .
8 ex:licCC-BY-NC  cc:permits       cc:DerivativeWorks .
9 ex:licCC-BY-NC  cc:requires      cc:Notice .
10 ex:licCC-BY-NC cc:requires      cc:Attribution .
11 ex:licCC-BY-NC cc:prohibits     cc:CommercialUse .

```

Listing 2.2: CC BY-NC 4.0 license represented in RDF (Notation 3 syntax), using CC REL vocabulary.

<sup>2</sup><https://creativecommons.org/licenses/by-nc/4.0/>

<sup>3</sup><https://www.w3.org/TeamSubmission/n3/>



## ODRL

The Open Digital Rights Language (ODRL) is a language for expressing rights and obligations over digital contents [29]. ODRL was initially introduced in 2000 and became a W3C Community Group in 2011. Since 2018, ODRL specifications are endorsed as W3C recommendations. The ODRL information model contains the following classes:

- **Policy**: the central entity than contains *Permissions*, *Prohibitions*, and *Duties*;
- **Asset**: a resource or a collection of resources;
- **Action**: the operation relating to an *Asset*;
- **Rule**: an abstract common ancestor to *Permission*, *Prohibition* and *Duty* classes;
- **Party**: an entity that undertakes Roles in a *Rule*;
- **Constraint**: an expression that puts a constraint on an *Action*.

Listing 2.3 expresses the same rights as the CC license reported above using ODRL.

```

1 @prefix odrl: <http://www.w3.org/ns/odrl/2/> .
2 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3 @prefix ex: <http://www.example.org/> .
4
5 ex:licCC-BY-NC      rdf:type          odrl:Policy .
6 ex:licCC-BY-NC      odrl:permission  _:Permission1 .
7 _:Permission1      odrl:action      odrl:reproduce .
8 ex:licCC-BY-NC      odrl:permission  _:Permission2 .
9 _:Permission2      odrl:action      odrl:distribute .
10 ex:licCC-BY-NC     odrl:permission  _:Permission3 .
11 _:Permission3     odrl:action      odrl:derive .
12 ex:licCC-BY-NC     odrl:prohibition  _:Prohibition1 .
13 _:Prohibition1    odrl:action      odrl:commercialize .
14 ex:licCC-BY-NC     odrl:duty         _:Duty1 .
15 _:Duty1           odrl:action      odrl:attribute .
16 ex:licCC-BY-NC     odrl:duty         _:Duty2 .
17 _:Duty2           odrl:action      odrl:attachPolicy .

```

Listing 2.3: CC BY-NC 4.0 license represented in RDF (Notation 3 syntax), using ODRL vocabulary.

## MPEG-21

In 2003, the Moving Picture Experts Group (MPEG), that covers most multimedia content subjects, produced MPEG-21 standard. MPEG-21 is an XML-based language that offers to declare rights and permissions based on the Rights Data Dictionary (RDD) [30]. The main element in MPEG-21 is the *License* that can have one or more *Grant(s)* and a license *Issuer*. Each *Grant* must include information about four elements:

- **Principle**: an entity (person, organization or device) to whom the rights are granted;

```

1 <?xml version="1.0" encoding="UTF-8" ?>
2 <r:license xmlns:r="urn:mpeg:mpeg21:2003:01-REL-R-NS"
3   xmlns:sx="urn:mpeg:mpeg21:2003:01-REL-SX-NS"
4   xmlns:mx="urn:mpeg:mpeg21:2003:01-REL-MX-NS"
5   xmlns:dsig="http://www.w3.org/2000/09/xmldsig#"
6   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
7   <r:grant>
8     <r:keyHolder>
9       <r:info>
10        <dsig:KeyName>Alice</dsig:KeyName>
11      </r:info>
12    </r:keyHolder>
13    <mx:print />
14    <r:digitalResource>
15      <r:nonSecureIndirect URI="http://www.example.org/rossi-0001"/>
16    </r:digitalResource>
17    <r:allConditions>
18      <sx:exerciseLimit>
19        <sx:count>3</sx:count>
20      </sx:exerciseLimit>
21    </r:allConditions>
22  </r:grant>
23 </r:license>

```

Listing 2.4: A sample license represented in XML, using MPEG-21 vocabulary.

- **Right:** an activity or an action (play, print, delete, etc.) that a **Principle** can exercise against some **Resource**;
- **Resource:** identifies an object to which the right in the *Grant* applies;
- **Condition:** one or more condition(s) that must be met before the right can be exercised.

Since MPEG-21 RDD contains only 14 rights, it is not possible to reproduce the above example. Therefore, we bring a simple license from [31] to close this Section. The subject in listing 2.4 is “Alice”, the object is “book”, the right is “print”, and the condition is “3 times”.

### 2.3.2 Policy Languages

As the Web became the main medium for selling products and services, enterprises try to apply automated techniques to analyze end-users’ personal information. Policy languages are designed to assure the end-users that their personal data are kept secure. In this section, we briefly introduce the two most established policy languages in the community.

### Platform for Privacy Preferences Project (P3P)

The P3P specification became a W3C Recommendation in April 2002 [32]. It is a form of ODRL for defining privacy preference in XML format. P3P serves two main goals: it allows websites to express their data collection practices in a standardized and machine-readable format; and it assists end-users to understand what kind of data will be collected by the websites they visit, how that data will be used, and what data/uses they may “opt-out” of or “opt-in” to.

In order for a website to use P3P, they have to place P3P policies on all their pages. On the other hand, the end-users can set their privacy preferences in a P3P built-in Web browser. In this case, when a user visits the target website, P3P compares what personal information the user is willing to release, and the information the server wants to get. If the two do not match, P3P will inform the user and asks if they are willing to proceed. For example, if the user visits a website with a pre-defined preference on “no collection of browsing history” and the website’s policy is set to using cookies, the browser will reject the cookie automatically. Some elements of a P3P policy are:

- **Access**: specifies whether the site provides access to various kinds of information. Some values for this element are: **none** (no access to identified data is given), **all** (access is given to all identified data), **contact-and-other** (access is given to identified online and physical contact information as well as to certain other identified data);
- **Categories**: Specifies the type of information stored in the cookie or linked to by the cookie. Some values are: **physical, financial, demographic, location**;
- **Purpose**: the purpose of data processing, e.g., **current** (Information can be used by the service provider to complete the activity for which it was provided), **develop** (to enhance, evaluate, or otherwise review the site, service, product, or market), **tailoring** (for tailoring or modifying the content or design of the site), **telemarketing** (to contact the data subject via a voice telephone call for promotion of a product or service);
- **Remedies**: Specifies the types of remedies in case a policy breach occurs. The three values are: **correct** (errors in connection with the privacy policy will be remedied by the service), **money** (in case of violation by the service provider, the data subject will be paid an amount specified in the human-readable privacy policy), **law** (remedies for breaches will be determined based on the law).

A small company should be able to deploy P3P in a few hours. As long as they have a clear understanding of the kinds of data their site collects and how the data are used, it is possible to use one of the P3P policy generator tools to easily create a P3P policy without having to learn XML.

### Enterprise Privacy Authorization Language (EPAL)

Enterprise Privacy Authorization Language (EPAL) is a formal language that enables enterprises to express their privacy policy [33]. Its core is an authorization scheme that specifies whether certain actions are allowed or not. The syntax of the set of rules for expressing a privacy policy will be different depending on the language used. However, the common elements in the policy are:

- **Data Users**: to classify individuals who have access to the data within an enterprise, e.g., *physician, nurse, etc.*;
- **Actions**: certain actions that are allowed regarding the data;

- **Data Categories:** defines the type of data which the company retains, e.g., *customer contact information, medical record, etc.*;
- **Purposes:** specifies the goal of collecting/using data, e.g., *customer order processing, marketing* and many more.

It should be clarified here that though P3P is an excellent language for expressing high-level privacy notices on websites, it is not suitable for formalizing an internal enforceable privacy policy. On the other hand, EPAL is explicitly designed to express an enforceable privacy policy within an enterprise.

## 2.4 Deep Learning Foundations

Deep learning is a subset of machine learning based on artificial neural networks. The word 'deep' comes from the use of multiple layers in the network. A simple neural network is a feedforward network that passes the data from one side (input layer) to the other side (output layer). In this Section, first, we explain how a simple neural network functions, and then a convolutional neural network is presented, which is the core of our machine learning solution throughout this thesis.

### 2.4.1 Multilayer Neural Network

A multilayer neural network is a type of feedforward artificial neural network that has an input layer, one or several hidden layers, and an output layer. The word 'hidden' comes from the fact that the machine has control over those layers. Figure 2.2 shows a fully connected multilayers neural network. Every neuron is connected to the subsequent layers of neurons in full, e.g., every orange line in the figure has a unique weight. Furthermore, every neuron has a unique bias. A neural network has its input data in the first layer. The input data are features from a single sample. For example, based on some sensor data (heat sensor, humidity sensor, ect.), we want to predict system failure or not failure. The network predicts the final output based on some given labels (supervised learning). So depending on the output neuron and the one with a higher value, the prediction is achieved.

The data gets passed through all hidden layers, and finally, it will be passed to the output layer. The neural network will randomly initialize the weights. The process of tuning all weights and biases is the actual training. Based on the labeled data, a loss value is calculated. The loss is a measure of how wrong the model is. Through the backpropagation, an optimizer adjusts the weights and biases in such a way that lowers the loss, slowly over time (learning rate). The learning rate, in part, dictates the size of the step that the optimizer takes to get to the best place. Since it is possible for the network to calculate and determine what weights it needs for loss to be zero and this will lead to overfitting<sup>4</sup>, a learning rate should be specified. In other words, the learning rate forces the model to learn the general principles.

As Figure 2.2 and equation 2.1 show, the information coming through from every unique input and every unique neuron has a unique weight associated with it, and they get summed per neuron. Bias is utilized to offset the values as opposed to the weight that changes the magnitude. Afterwards, the summed information runs through an activation function which is calculated for every single layer (equations 2.2, 2.3). Activation functions determine that final output before it becomes an input to another layer or the final output of the network. They decide to what degree a neuron is fired (if fired at all). One of the simple activation functions is the step function which outputs one if  $x > 0$  and zero if  $x \leq 0$ . The problem with the step function is that when the loss function is being calculated, and optimizer is trying

---

<sup>4</sup>The production of an analysis that corresponds too closely or exactly to a particular set of data and may, therefore, fail to fit additional data or predict future observations reliably (Definition of "overfitting" at OxfordDictionaries.com for statistics).

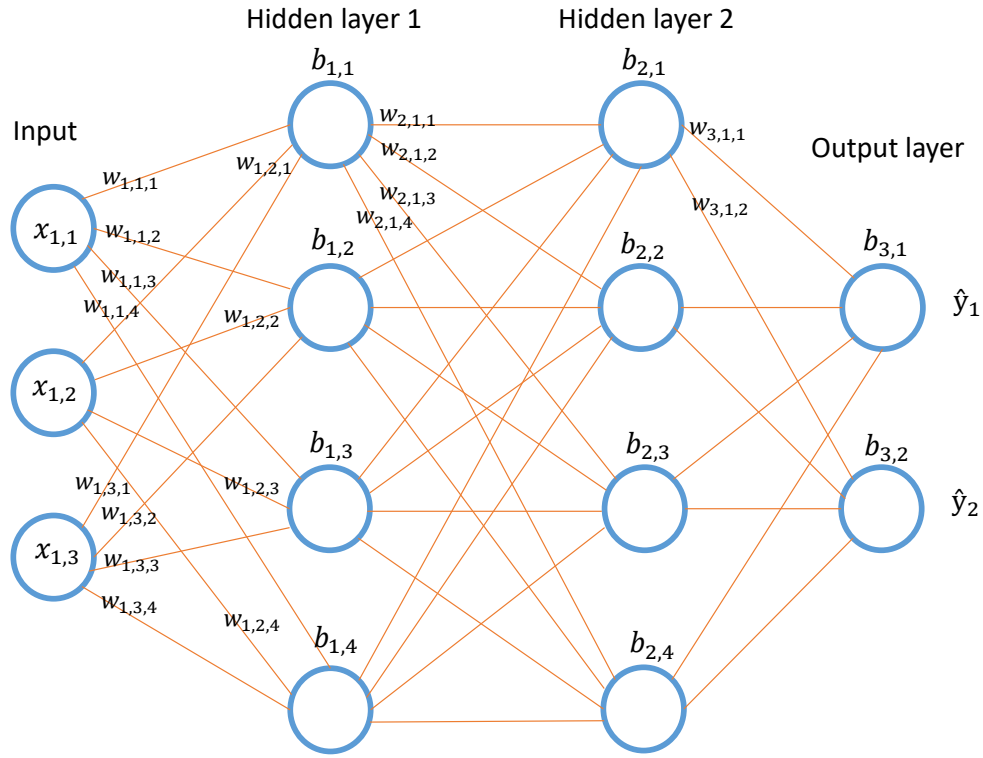


Figure 2.2: A fully connected feedforward multilayers neural network.

to tune the weights, there is no granularity to determine how close the neurons were to fit into the labeled data. Therefore, conventionally ReLU is used as the activation function.

$$ReLU(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

For the output layer, the softmax or sigmoid functions are used to get distribution. Equations 2.4, 2.5 and 2.5 show the calculations for the output layer. After a full forward pass, the loss will be measured. One of the common loss functions for the classification problem is the cross-entropy loss (equation 2.7). The cross-entropy loss is a negative log loss due to the nature of neural networks. Finally, combining all the above equations, we reach the formula shown in 2.8. In this equation,  $X$  is the input to the neurons, and they could be either from the true input layers, or they could be outputs from neurons in the hidden layers.

$$z_1 = \sum_{i=1}^{n_0} x_{1,i} w_{1,i,j} + b_{1,j} \quad (2.1)$$

$$y_1 = ReLU(z_1) = \max(0, z_1) \quad (2.2)$$

$$y_1 = \mathcal{V}_{j=1}^{n_1} \max(0, \sum_{i=1}^{n_0} x_{1,i} w_{1,i,j} + b_{1,j}), y_2 = \mathcal{V}_{j=1}^{n_2} \max(0, \sum_{i=1}^{n_1} x_{2,i} w_{2,i,j} + b_{2,j}) \quad (2.3)$$

$$z_{3,j} = \sum_{i=1}^{n_2} x_{3,i} w_{3,i,j} + b_{3,j} \quad (2.4)$$

$$\hat{y}_1 = \text{softmax}_1(z_3) = \frac{e^{z_{3,1}}}{\sum_{k=1}^{n_3} e^{z_{3,k}}}, \hat{y}_2 = \text{softmax}_2(z_3) = \frac{e^{z_{3,2}}}{\sum_{k=1}^{n_3} e^{z_{3,k}}} \quad (2.5)$$

$$\hat{y} = \mathcal{V}_{j=1}^{n_3} \frac{e^{\sum_{i=1}^{n_2} x_{3,i} w_{3,i,j} + b_{3,j}}}{\sum_{k=1}^{n_3} e^{\sum_{i=1}^{n_2} x_{3,i} w_{3,i,k} + b_{3,k}}} \quad (2.6)$$

$$L = - \sum_{l=1}^N y_l \log(\hat{y}_l) \quad (2.7)$$

$$L = - \sum_{l=1}^N y_l \log\left(\mathcal{V}_{j=1}^{n_3} \frac{e^{\sum_{i=1}^{n_2} (\mathcal{V}_{j=1}^{n_2} \max(0, \sum_{i=1}^{n_1} (\mathcal{V}_{j=1}^{n_1} \max(0, \sum_{i=1}^{n_0} X_i w_{1,i,j} + b_{1,j}))_i w_{2,i,j} + b_{2,j}))_i w_{3,i,j} + b_{3,j}}}{\sum_{k=1}^{n_3} e^{\sum_{i=1}^{n_2} (\mathcal{V}_{j=1}^{n_2} \max(0, \sum_{i=1}^{n_1} (\mathcal{V}_{j=1}^{n_1} \max(0, \sum_{i=1}^{n_0} X_i w_{1,i,j} + b_{1,k}))_i w_{2,i,j} + b_{2,k}))_i w_{3,i,k} + b_{3,k}}}\right) \quad (2.8)$$

## 2.4.2 Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of neural network with several layers of convolutions. A convolution is an  $m \times n$  window that slides over the input matrix (also called  $m$  by  $n$  convolution kernel). As opposed to the feedforward network that connects each input neuron to each output neuron in the next layer, CNN uses convolutions to compute the output. Moreover, unlike the traditional neural network they do not need the input data to be flattened, so it is possible to pass a  $k$ -dimensional input. The goal of convolution is to locate the features of the input matrix. Traditionally, CNNs were used for image processing, because every image is a two-dimensional matrix with image pixels. However, with the emergence of word embeddings in Natural Language Processing (NLP), CNNs have been extensively applied to NLP problems.

Word embeddings are multi-dimensional vectors that represent a word in a vector space. In order to build word embeddings, a large corpus of correct text is required. The algorithms adapt unsupervised machine learning to build a model using a large corpus. After the model creation, an input word (or input paragraph/document) in a natural language text will be converted to a large-dimensional vector space and then using one of the vector similarity measures (Jaccard, Dice, Cosine, Euclidean, Manhattan, etc.), the closet vectors and their corresponding texts will be found. Figure 2.3 shows a widely-used cosine similarity measure in a two-dimensional space. According to this figure, the semantic similarity between “Burger” and “Sandwich” is 0.6, which will be interpreted as very similar (maximum = 1).

Representing texts as multi-dimensional vectors makes it feasible to apply CNNs on top of textual resources. Figure 2.4 shows a CNN from [9]. Instead of image pixels, the input to most NLP tasks is sentences or documents represented as a matrix (in this case, 5-dimensional vector). There are three different convolutions:  $4 \times 5$ ,  $3 \times 5$ , and  $2 \times 5$ . Each convolution has two filters that are initialized randomly. The first layer of convolution tends to find very basic features of the input matrix, and then it will pass the detected features to the next layer, which sees those basic features and try to find more complex features. The CNN keeps sliding the window over the entire matrix, so basically condensing that matrix. After the

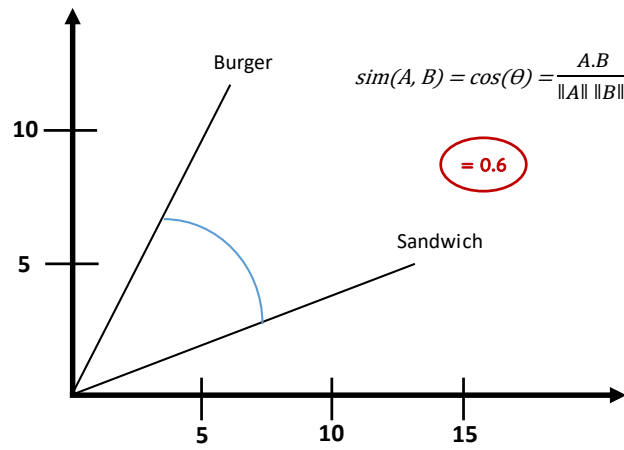


Figure 2.3: Cosine similarity.

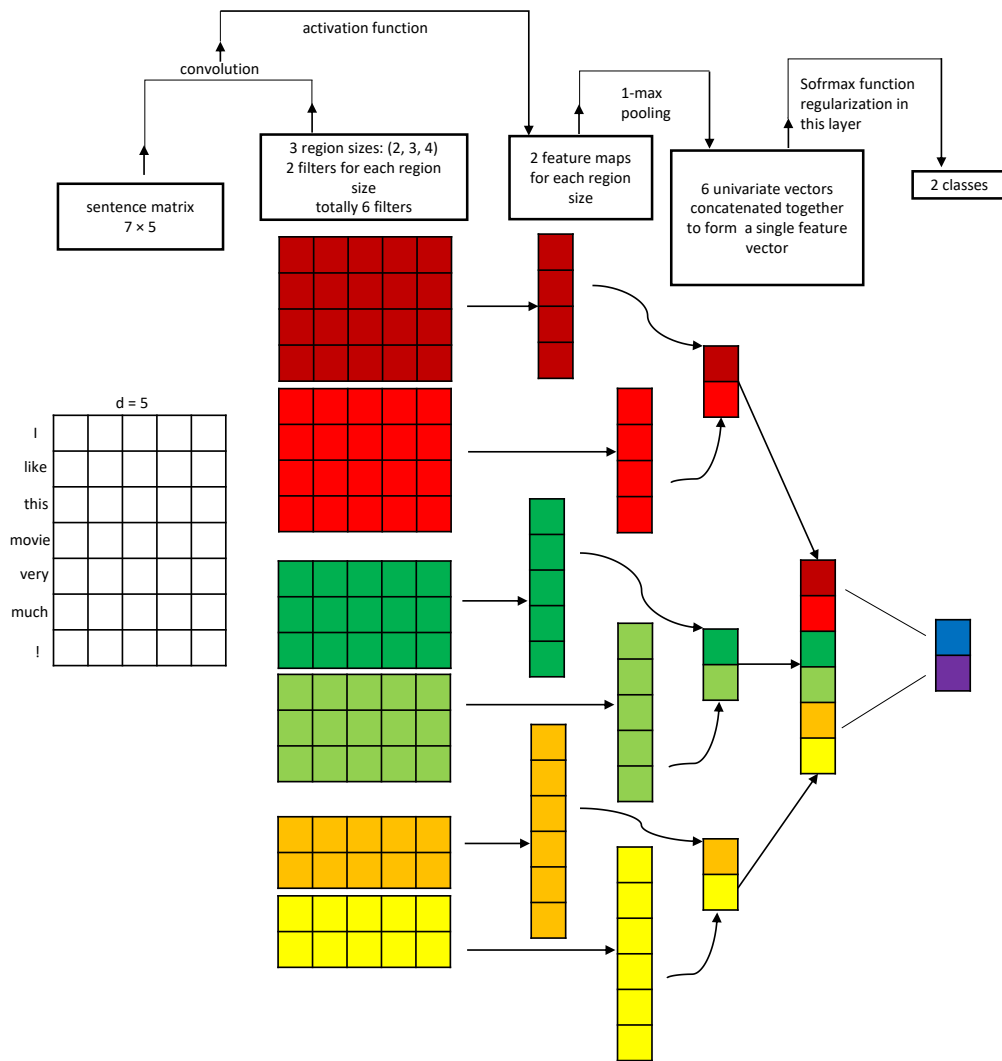


Figure 2.4: A Convolutional Neural Network (CNN) architecture for sentence classification [9].

completion of each convolution, a pooling is applied. Once again, the pooling has a window, and the most common form of pooling is max-pooling, which takes the maximum value in that convolution window. Generally speaking, CNN reduces the input matrix to the basic building blocks, and find patterns of those blocks given how many layers it has. We refer interested readers to [34] for further elaborations on how CNNs are used for NLP problems.

## 2.5 Background Overview

The research problem of extracting valuable information from contractual agreements require a comprehensive approach from different angles. The concepts and foundations presented in this Chapter lay a solid foundation for addressing the posed challenges. The main use-cases of this thesis are end-user license agreements and privacy policies, which were introduced in Sections 2.1.1 and 2.1.2. The ODRL vocabulary described in Section 2.3.1, serves as our main vocabulary for employing vocabulary-based information extraction in Chapter 4. Chapter 5 reproduces a state-of-the-art result using CNN. In addition, Chapters 5 and 6 specifically present approaches for the structured presentation of privacy policies to facilitate end-user interaction.



---

## Related Work

---

This Chapter reviews community efforts and state-of-the-art approaches related to the main research problem and research questions. In Section 3.1, we analyze methodologies and techniques for the enrichment of legal text, using semantic annotations. Our literature review suggests that semi-automatic annotations tailored for contractual agreements have not yet gained a necessary momentum. This motivates our research in *semantic interpretation of contractual agreements using ontologies*. Section 3.2 describes advances in machine learning approaches for contractual agreements analysis and interpretation. We observe that, despite encouraging results gained by previous studies, there is a lack of a standard benchmark, which makes it difficult to interpret and compare different results collectively. Section 3.3 studies the feasibility of detecting unlawful clauses in enterprise policies. Here we observe a number of attempts by the literature to determine regulatory compliance of human-readable agreements; further motivating our research.

### 3.1 Legal Texts Enrichment using Semantic Annotations

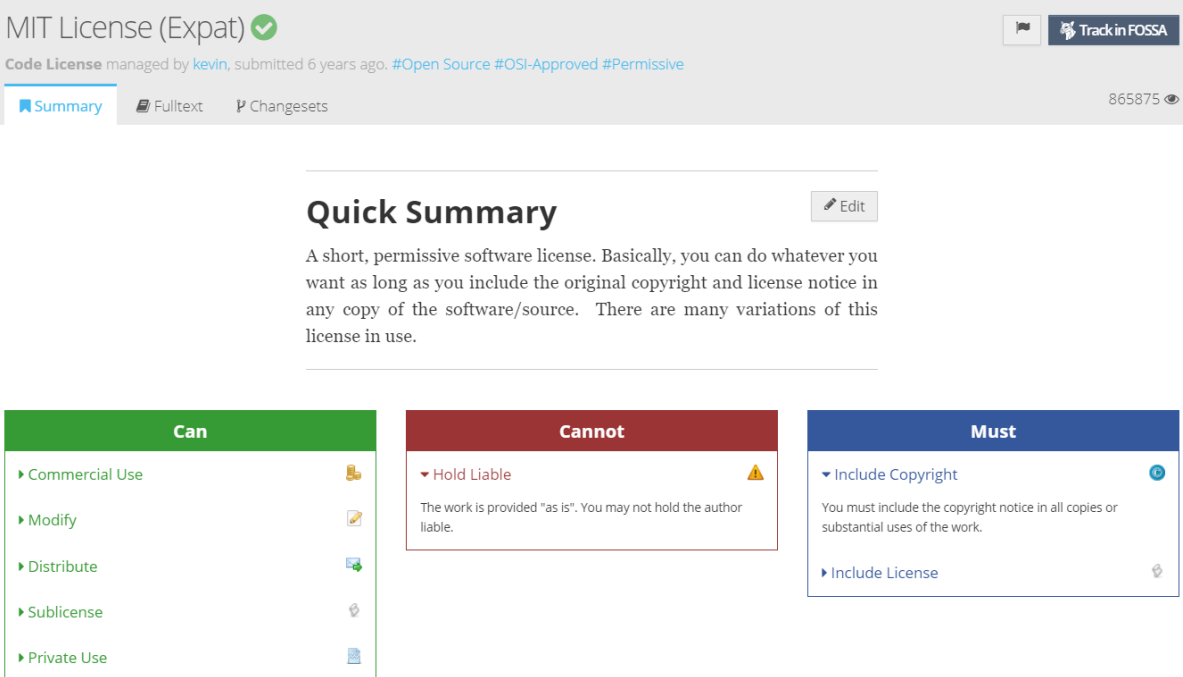
Natural Language Processing (NLP) techniques fill the gap between legal texts and their semantic level and accelerate the time-consuming annotation process. Several NLP tools allow the extraction of the legal knowledge embedded in the legal text and permit to express the extracted information in machine-processable languages and structures. In the last twenty years, various legal XML standards such as RuleML, RIF, and SWRL were defined and created to represent legal text and rules [35]. In addition, with the advance of Semantic Web, legal ontology research incorporated with NLP extraction of semantics, significantly improved the legal concepts modeling [36–39]. On the advanced level, semantic technologies are applied for regulatory compliance change management. Enriching regulations with regulatory ontologies enable the executives and legal practitioners of an enterprise, to identify and extract compliance obligations and prohibitions. In this Section, we study manual and semi-automatic approaches for semantic tagging of lengthy and complex legal texts.

#### 3.1.1 Manual Enrichment

Due to the complexity of contractual agreements and given that many people ignore them, there have been a few initiatives that use collaborative approaches in order to give a brief summary of consent forms. *Tldrlegal*<sup>1</sup> is an online service, which uses a manual, crowdsourced way to help people understand the most commonly-used licenses. It is supported by users, and everyone can create an account and suggest

---

<sup>1</sup><http://tldrlegal.com>



MIT License (Expat) ✓

Code License managed by kevin, submitted 6 years ago. #Open Source #OSI-Approved #Permissive

Summary Fulltext Changesets 865875

### Quick Summary Edit

A short, permissive software license. Basically, you can do whatever you want as long as you include the original copyright and license notice in any copy of the software/source. There are many variations of this license in use.

Can	Cannot	Must
<ul style="list-style-type: none"> <li>Commercial Use</li> <li>Modify</li> <li>Distribute</li> <li>Sublicense</li> <li>Private Use</li> </ul>	<ul style="list-style-type: none"> <li>Hold Liable</li> </ul> <p>The work is provided "as is". You may not hold the author liable.</p>	<ul style="list-style-type: none"> <li>Include Copyright</li> <li>Include License</li> </ul> <p>You must include the copyright notice in all copies or substantial uses of the work.</p>

Figure 3.1: Tldrlegal service showing MIT license summary.

a short summary of a chosen license. Figure 3.1 shows a screenshot from the *tldrlegal* website. In addition to the summary, the full text and the changing history of the license is available. Furthermore, a green checkmark icon (next to the license name), confirms that the summary was verified by the service provider’s legal experts. Likewise, *Data Right Finder*<sup>2</sup> is another service that provides an analysis of financial services’ privacy policies, based on the GDPR. It is an open service, and anyone can contribute, use and reuse it under the Open Database License<sup>3</sup>. The team also publishes the machine-readable privacy notices in their Github page<sup>4</sup>. Figure 3.2 presents a partial snapshot from *Data Right Finder* service where PayPal’s privacy notice is manually analyzed according to the GDPR.

In addition to the crowdsourcing approach, the manual annotation of legal texts, in general, can be applied either on the whole text or a part of the text as a training set to be used in the automatic algorithms. Conventionally, the Subject Matter Experts (SMEs) annotate the text using one of the two schemes:

- Goal-driven approach: a set of policy-specific goals are defined which guide the process of extracting goal statements from regulations;
- Ontology-based approach: SMEs identify regulation concepts from vocabularies or ontologies, which include domain concepts.

A set of standards that are widely used for regulation and policy interpretation are the specifications provided by the Object Management Group (OMG). OMG is an international consortium whose mission is developing technology standards for industries. In the sequel, we provide a brief overview of the most common OMG specifications.

<sup>2</sup><https://www.datarightsfinder.org/>

<sup>3</sup><https://opendatacommons.org/licenses/odbl/summary/index.html>

<sup>4</sup><https://github.com/datarightsfinder/data>

The screenshot shows the 'Data Rights Finder' interface. At the top, it says 'Data Rights Finder' and has links for 'About', 'Developers', and 'Contribute'. The main heading is 'PayPal'. Below it, a sub-heading reads 'We read PayPal's privacy policy so you don't have to.' followed by two bullet points: 'Find out what they do with data about you' and 'Contact them if you have a request about that data'. A prominent orange button says 'Make a data request'. On the left, a 'Contents' section lists various categories like 'Organisation information', 'Data Protection Officer', 'Data categories collected', etc. The main content area is divided into sections: 'Data categories collected' (explaining that organisations must give details about what categories of data are stored and processed, and listing categories like Bank transactions, Credit history, Device information, Email address, Location, Names, Postal address, and Telephone number), 'Unusual processing purposes' (explaining that organisations must provide information about what they do with data, and noting that this privacy notice does not appear to mention any unusual processing purposes), and 'Third parties' (explaining that organisations must give details about other parties that personal data is shared with).

Figure 3.2: Data Rights Finder service showing analysis of PayPal privacy policy.

### OMG Specifications

OMG standards encompass several industry sectors, e.g., finance, government, healthcare, manufacturing, military, retail, and space exploration. An established specification in business modeling is the Semantics Of Business Vocabulary And Rules (SBVR) [40]. SBVR is an ISO terminological dictionary for defining business concepts and rules, represented in simplified natural language. It contains noun concepts, verb concepts, and rules (definitional and behavioral) for the business domain. The SBVR vocabulary comprises:

- **General Noun concepts**, which correspond to classes of object in business domain, e.g., *Share*, *Bank*;
- **Individual Noun concepts**, which correspond to individual occurrences in business, e.g., *Deutsche Bank*;
- **Verb concepts**, which correspond to relationships between noun concepts (either general or individual), e.g., *Bank transfer Share*.

The SBVR vocabulary is widely used to express policies and rules based on obligation, possibility, and prohibition. Every rule is a combination of 1) a modality; 2) one or multiple verb concepts connected with keywords. Let's consider the following rule:

An Obligation Example

It is obligatory that each Price reflects the Prevailing Market Condition for each Share .

- The modality of this rule is expressed in “It is obligatory that” which indicates an obligation.
- There are three **Noun Concepts**: Price, Prevailing Market Condition, and Share.
- We have two **Verb Concept**: Price reflects Prevailing Market Condition and Share has Prevailing Market Condition.

Following the above convention, SMEs are able to interpret a legal text based on the SBVR specification. Listing 3.1 shows an example of interpretation protocol from [41]. Applying this protocol, will convert the provision “Money services business must establish procedures to verify the identity of a person who obtains prepaid access under a prepaid program”<sup>5</sup> to the following rule:

“It is obligatory that each money services business **establishes** procedures to **verify** the identity of the person **obtaining** prepaid access under a prepaid programme.”

- 1 1. Read the text.
- 2 2. Identify the modalities (obligations, prohibitions, etc.).
- 3 3. For each modality:
  - 4 a. Add the relevant modality keyword;
  - 5 b. Identify the (English) verb on which the modality is applied;
  - 6 c. Stylise this verb using the SBVR verb style (binary,
  - 7 unary, general, ...);
  - 8 d. Identify the noun concepts (general, individual, etc.) or
  - 9 the verbal phrase(s) playing the roles in this verb;
  - 10 e. If the verb roles are played by noun concepts, complete
  - 11 the SBVR modified verb concept by stylising the identified
  - 12 noun concepts;
    - 13 i. Add all the stylised noun concepts to the
    - 14 noun\_concepts\_list
    - 15 f. If the verb roles are played by verbal phrases, stylise
    - 16 each verbal phrase by identifying English verbs, SBVR noun
    - 17 concepts and keywords;
      - 18 i. Add each verbal phrase to the
      - 19 supporting\_verb\_concepts\_list
  - 20 4. For each noun concept in noun\_concepts\_list:
    - 21 a. Start enriching by identifying the characteristic of each
    - 22 noun concept, if any (e.g., necessary characteristic);
    - 23 b. Identify other definition elements.

Listing 3.1: SBVR-based rule interpretation protocol.

<sup>5</sup>The provision is from the “anti-money laundering programs for money services businesses” regulation.

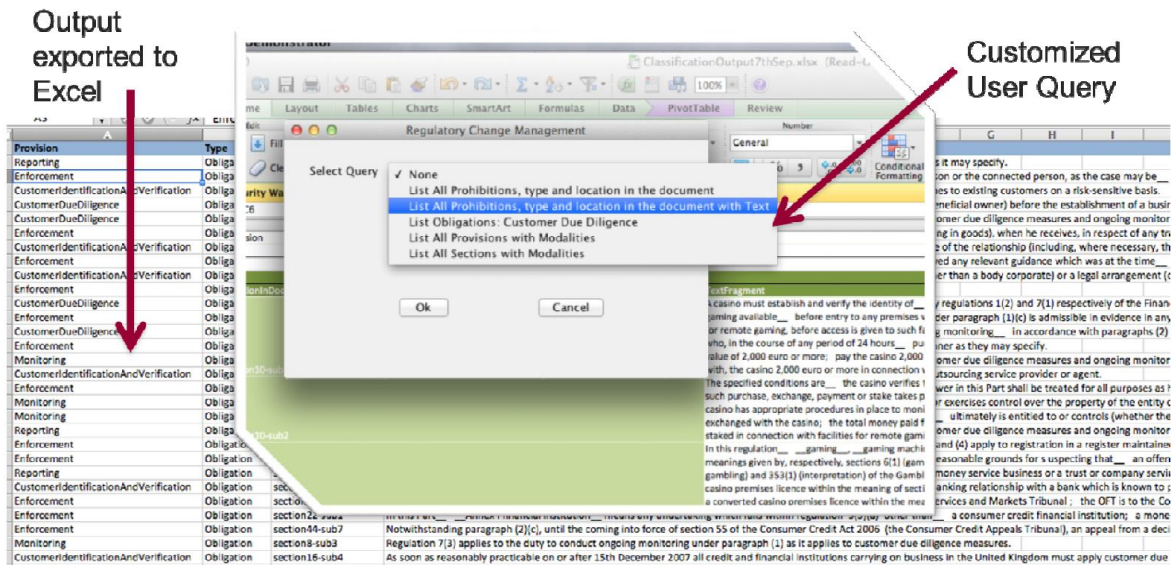


Figure 3.3: GRCTC's developed tool for querying the semantically enriched text.

SMEs use different tools for interpreting legal texts. For instance, *Designs for Management*<sup>6</sup> is an SBVR editing suite that validates SBVR interpretations and generates machine-readable vocabularies and rules in the XML format based on the SBVR metamodel. The XML interpretation of the rules allows further to apply semi-automatic conversion of SBVR XML to OWL<sup>7</sup>. In this case, the SMEs can send their desired queries. The query will be translated to a SPARQL query (a semantic query language for knowledge bases) [42], and the structured result will be presented to the user. Figure 3.3 shows a snapshot of a sample query implemented by [43]. In this study, a part of Anti-Money Laundering regulation is manually annotated by SMEs. Then the manual annotations are used as a training set for the automatic classification algorithms. Finally, the trained model is applied to the un-annotated parts of the text to automatically tag them with the concepts from the ontology.

Beside SBVR, OMG's FIBO (Financial Industry Business Ontology) standards, such as FIBO foundation and FIBO business entities, are also used in business regulation interpretation.

### Goal-driven Approaches

In requirement engineering, goals are considered as the essential component involved in the process. According to [44], "Goal-oriented requirements engineering is concerned with the use of goals for eliciting, elaborating, structuring, specifying, analyzing, negotiating, documenting, and modifying requirements". In this sense, in goal-driven text annotation, first, a set of domain-specific goals are defined, and then the goal-mining heuristics guide the process of extracting goal statements from the target text.

In [45], a set of goals in the privacy policies were identified by answering two questions:

- What goal(s) does this statement or fragment exemplify?
- What goal(s) does this statement obstruct?

<sup>6</sup><https://designsformanagement.com/>

<sup>7</sup>Ontology Web Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge.

Based on the above questions, important keywords and actions from privacy policies were documented in a Web-based Privacy Goal Management Tool (PGMT). An example goal in the repository would be: “PROVIDE access to CI (Customer Information) to authorized personnel with authorized roles”. Each goal in the repository is associated with a unique ID, a description, and an actor. In total, 1 032 goals were extracted from 40 privacy policies.

Using the PGMT goals, [46] and [47] developed semantic models from privacy policy goals mined from policy documents. The process of developing semantic models includes two main stages: semantic parameterization and goal analysis. The semantic parameterization has three steps:

1. **Restating goals into RNLs (Restricted Natural Language):** RNL has exactly one primary actor, action, and at least one object (as opposed to the PGMT goals that may describe nested activities). For instance, in the PGMT goal G161, COLLECT information from nonaffiliates, the action is identified as “collect”, the object is “information”. The actor is already identified as the “provider” in the PGMT;
2. **Building semantic models:** after identifying the essential elements of a component (e.g., actor, action, and object), the semantic models are built by assigning words from a RNL with a part-of-speech tag, e.g., <provider (noun) > <collect (verb) > <information (noun) >. This process is called parameterization;
3. **Formalization in a context-free grammar:** after the parameterization process is completed, the semantic models are described by a context-free grammar (CFG) and is supported by a qualitative and quantitative policy analysis tool. Expressing the semantic models in CFGs ensures the correctness of parameterization process.

The studies show through a few examples that applying semantic models enables policy statements and comparison identification of potential limitations that exist in policy languages.

Last but not least, following the goal-oriented scheme, [48] derives security requirements from regulations to support the software engineering effort. A methodology is presented for extracting access rights and obligations from regulation texts on the statement-level. The methodology identifies six types of data access constraints: *subject, action, modality, object, target, purpose*. It also handles complex cross-references, resolves ambiguities, and assigns priorities between access rights and obligations to ensure regulatory compliance. In order to implement patterns for *basic activity pattern with modality, purposes, nouns distinguished by verb phrases*, and *rules or conditions*, the efforts by the previous mentioned studies were used [46, 47]. The methodology was applied on the entire regulation text of the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. First, two persons working in tandem extracted rules from the text that required close to 26 hours. Then the priorities were extracted that required close to 29 hours. In total, 300 access rules were extracted, which are comprised of 1 894 constraints. 50 rules out of were refrainments (deny access), and among 58 extracted exceptions, there were more than 12 205 priorities between different rules.

All the above efforts, presented here, require intensive time and energy. Since these efforts depend heavily on manual work, they are barely extendable to other rules and regulations. Furthermore, although OMG specifications are created and refined by the experts, the goal-driven approaches infrequently involve domain experts in the annotation process, which in the field of legal text interpretation, is a major limitation.



### 3.1.2 Phrase Heuristics (Linguistic Rules)

In the previous Section, we studied those text annotation approaches, that are heavily based on the manual tagging. In this Section, we briefly address semi-automatic legal text enrichment. Semi-automatic text annotations use linguistic rules to annotate the text, and since these rules are hand-coded and are implemented by humans, the whole process is considered as “semi-automatic”. In the remaining subsections, first, we explain common terminologies in legal texts, and then we specifically review those efforts, that their rules definition are based on ontologies and vocabularies.

#### Common Terminologies in Legal Text

Legal documents are written in a structured and clear format. As opposed to short texts (such as Tweets), they have clear, though complicated terminologies. This attribute allows for rule-based interpretation. The common terms in legal documents are:

- **Stakeholder or Entity:** an individual, natural person or a legal entity exercising rights under, and complying with all of the terms of the specific regulations or policies;
- **Right or Permission:** a statement about one or more actions that an entity or a stakeholder is permitted to perform;
- **Refrainment or Prohibition:** a statement about one or more actions that an entity or a stakeholder is NOT permitted to perform;
- **Obligation or Duty:** a statement about one or more actions that an entity or a stakeholder is required to perform;
- **Constraint Phrase:** a statement that restricts the possible interpretations for a concept; For instance, the phrase “an end-user who has an account” constrains the set of all possible end-users;
- **Normative Phrase:** a statement that expresses a value judgment about whether a situation is permitted, desired, or prohibited. These statements are characterized by the modal verbs "should", "may" or "must", etc.;
- **Rule Statement:** a rule can be a right, obligation, or refrainment and may be restricted by constraints.

The following terms are the foundation for implementing linguistic rules, and they are closely related to each other. For example, if a healthcare patient has a right to access (permission) their health records, then their physician’s office has an obligation to provide access. The permissions, obligations, and refrainments are distinguished by the normative phrase, and the whole statement is called a rule.

The clear structure and legal language of regulations and policies have popularized a subset of IE, called Ontology-Based Information Extraction (OBIE). Here, domain ontologies guide defining the linguistic rules and are used by the information extraction process. In the next part, we provide a brief overview of OBIE approaches applied to the legal domain.

#### Ontology-Based Information Extraction

Ontology-Based Information Extraction (or OBIE) uses a formal ontology as one of the inputs for information extraction, and the linguistic rules are defined based on the ontology specification. ART [49]

is a framework that extracts meaningful entities from regulatory text. It exploits concepts in the process ontology to annotate the regulatory text. First, the regulatory text is pre-processed, and it is manually converted to HTML format. Second a *Feature Reader* identifies the document features such as font-style, font-weight, font-family, font-color and text-content. Third, based on the document features, a *Structure Predictor*, predicts the component of the document, e.g., the large font-size of a text means it has a higher level in the structure hierarchy (chapter, section, paragraph). Fourth, the predicted components are confirmed by a user via a user interface. After the semi-automatic generation of the document structure and identifying paragraphs, the regulation-entities such as subject, obligation, and action are extracted using normative phrases and the ontological concepts. The study assumes that the modal verbs such as "must" and "should" are indicators of a possible obligation (or refrainment depending on the existence of a negation). In addition, a subject in a statement is the words upon which the requirements and expectations are imposed, and the main verb of the statement is the action. Based on the above assumptions and using a dependency parser, the chunk boundaries, such as subject-chunk, obligation-chunk, action-chunk, and condition-chunk are identified. Each chunk is linked to the ontology, if possible. As a proof-of-concept, the approach was applied to one of the EU regulation in the pharmaceutical industry. A process-ontology for the pharmaceutical processes and a regulatory ontology were used from the existing research [37, 50]. The results were evaluated against the manually created annotation, and according to the authors, were very close to the gold standard.

GaiusT is a tool that semi-automatically extracts rights and obligations from a regulatory document in order to assist requirement engineers of an enterprise [51]. It uses semantic annotation techniques, where the legal text is annotated with the concepts of an ontology. Similar to ART, GaiusT leverages common terminologies in legal texts. First, a structural analysis identifies the hierarchical structure of the document, basic text constructs, and cross-references. Then a comprehensive semantic analysis detects normative elements of a document such as, *rights*, *obligations*, *anti-rights*, and *anti-obligations*. An existing framework [52] is adopted and extended for the semantic annotation, in which textual annotations for concepts are inferred based on a domain-specific ontology. GaiusT extends the previous framework by including new components specific to the annotation of legal documents in different languages (English & Italian). The Syntactic indicators for the deontic concepts are defined manually:

- *Right or Permission*: may, can, could, permit, to have a right to, should be able to;
- *Anti-Right*: does not have a right to;
- *Obligation*: must, requires, should, will, would, which is charged with, may not, can not, must not;
- *Anti-Obligation*: is not required, does not restrict, does not require.

The rules are defined according to the syntactic nature of legal documents in different languages. For instance, in the Italian regulation, a statement expressing an obligation normally uses the present active or present passive tense, (e.g., “the organization sends a request”), where, in English regulations, obligations are usually expressed with modal verbs such as “must” or “should”.

GaiusT was evaluated on two regulatory documents: HIPAA Privacy Act (US, in English) and the Stanca law (Italy, in both English and Italian). For the English resource, GaiusT was able to identify legal requirements with high precision (from 93 to 100%) and good recall (70 to 100%), apart from anti-obligation (33%). For the Italian regulatory document, the precision ranges from the 67% for *Actor* to the highest rate 100% for *Right*, *Anti-Obligation* and *Constraint*. However, recall was not as high as precision: 33% for *Right* and 35% for *Constraint*.

In a similar study, a process called *Semantic Parameterization* has been applied to privacy rules from HIPAA [53]. First, rights and obligations are reformulated into Restricted Natural Language Statements



(RNLS), and then RNLSs are mapped into semantic models to clarify ambiguities. In the end, the authors listed some limitation which arose from their incomplete set of phrase heuristics.

The efforts discussed in Section 3.1 are the foundation for our approach in Chapter 4. To the best of our knowledge, there is no OBIE methods applied to contractual agreements. The promising results gained by state-of-the-art convinced us to also pursue an ontology-based approach in this thesis.

## 3.2 Interpretation of Contractual Agreements using Machine Learning

Over the past years, the Machine Learning (ML) revolution has made significant advances in Natural Language Processing (NLP). Likewise, legal text processing as a sub-field of NLP has drawn the researchers' attention [54–57]. However, considering the abundance of previous ML studies in law and legal texts, in this Section, we focus on those efforts that specifically address contractual agreements and regulatory documents.

### 3.2.1 Linear Classification Methods

Linear classifiers are a subfield of machine learning and aim at using an object's features to identify which class it belongs to. Linear classification methods achieve this goal based on the value of a linear combination of the features. Below, we provide an overview of studies that exploited linear classification algorithm to analyze legal documents.

The *NLL2RDF* framework exploits machine learning and Support Vector Machine (SVM) to generate RDF expressions of license agreements [58], targeting open linked data as their primary use-case. The authors used ODRL and CC REL vocabulary to manually annotate the dataset and build a gold standard. Similarly, *NLL2RDF* also is primarily concerned with Permissions (derive, reproduce, modify, copy, sell), Prohibitions (commercialize), and Duties (shareAlike, attachPolicy, attribute). However, the framework's limitation is that it only covers a limited number of rights and conditions. Furthermore, notwithstanding that their dataset covered 37 licenses, the class with the highest frequency only scored 28 occurrences. This low number might be related to their training data, since after going through their publicly available dataset<sup>8</sup>, we noticed a scarce number of annotations in the 4-5 page licenses.

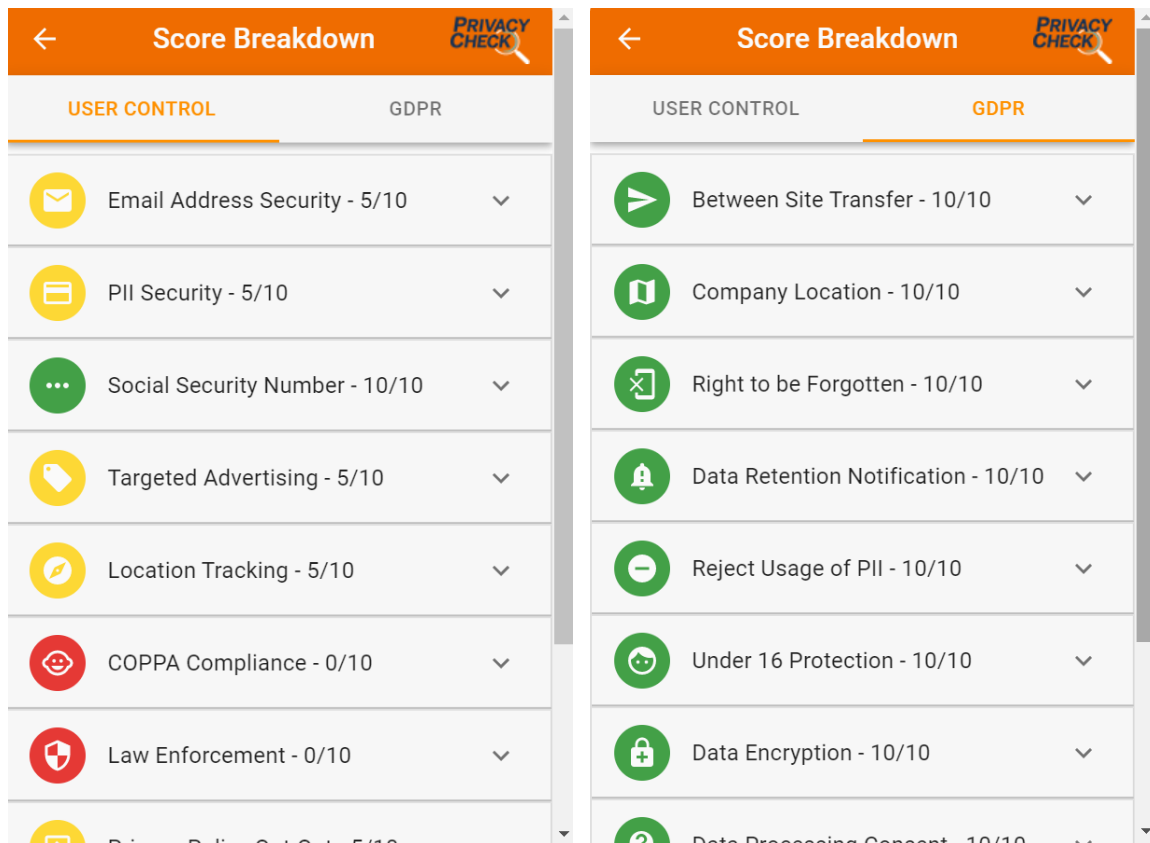
A prominent group on privacy policy analysis is *Usable Privacy Policy Project*<sup>9</sup>. They provided OPP-115, the first comprehensive dataset with fine-grained annotations on paragraph level [59]. The project aims to extract important information for the benefit of regular and expert end-users. To do so, a corpus containing 115 privacy policies from 115 US companies was annotated by three experts on the paragraph level (10 experts in total and three experts per document). Along with the creation of the dataset, the authors built different ML models for the prediction of high-level categories. The gold standard for evaluating the methods was compiled based on majority votes: if two or more experts agreed on a single category, it was considered in the final gold standard. The best-reported micro-average F1 is 66% that was achieved with Support Vector Machine (SVM).

A few approaches developed a model with supervised ML to measure completeness of privacy policies [60, 61]. The dataset used in training contains a set of pre-defined categories based on privacy regulations and guidelines. Finally, the trained model predicts a category for an unseen paragraph. Once again, none of the corpora were created with the full support of experts, which is an essential prerequisite in legal text processing.

---

<sup>8</sup><http://www.airpedia.org/nll2rdf/dataset-licenses/>

<sup>9</sup><https://usableprivacy.org/>



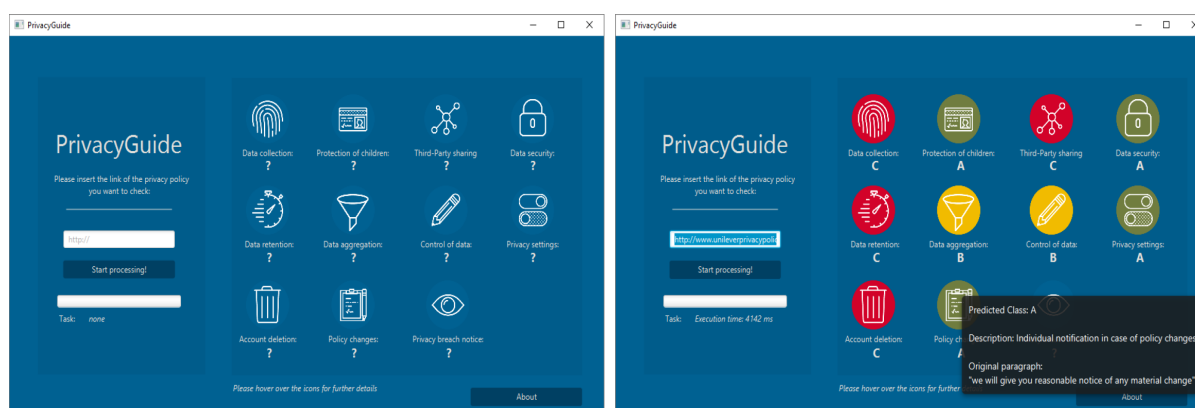
(a) The user control score, based on 10 privacy concerns.

(b) The GDPR compliance score.

Figure 3.4: Evaluation of ResearchGate privacy policy using the *Privacycheck* Chrome extension.

*PrivacyCheck* is an approach for automatic summarization of privacy policies using data mining [62]. It answers ten pre-defined questions concerning the privacy and security of users' data and is also available as a Chrome browser extension. Figure 3.4 illustrates the automatic analysis of *PrivacyCheck* Chrome plugin, applied to the *ResearchGate* privacy policy. In order to train the model, a corpus containing 400 privacy policies was compiled, and seven privacy experts manually assigned risk levels (Green, Yellow, Red) to the ten factors. First, a pre-processing step finds those paragraphs that have at least one keyword related to one of 10 factors. The methodology of selecting keywords was largely manual. Then, the selected paragraphs will be sent to a data mining server where 11 data mining models were trained, one for checking if the corresponding page is a privacy policy and one each for the ten questions. The authors claim that, on average, 60% of the times, *PrivacyCheck* finds the correct risk level. The limitation of *PrivacyCheck* is its lack of an Inter Annotator Agreement (IAA) for the annotators. According to the paper, the quality control was performed by assigning each policy to two team members. However, only 15% of privacy policies were compared, and their discrepancies were resolved, which makes the training dataset less reliable.

*PrivacyGuide* is another summarization tool inspired by GDPR that classifies a privacy policy into 11 categories using NLP and machine learning and further measures the associated risk level of each class [63]. Figure 3.5 shows a snapshot of *PrivacyGuide* automatic summarization applied to the *Unilever* privacy policy. The red icons illustrate a high-risk prediction for the corresponding category. Similar to previous studies, *PrivacyGuide* uses the three-level scale risk based on classification (i.e., Green, Yellow,



(a) The welcome screen.

(b) After applying a sample policy.

Figure 3.5: PrivacyGuide snapshots.

Red). The 11 criteria and their associated risk levels were defined by GDPR experts. Based on these criteria, a privacy corpus was compiled with the help of 35 university students. Each participant assigned a privacy category to text snippets and classified them with a risk level. The author reported that the weighted average accuracy is 74% for classifying a privacy policy into one of the 11 classes, and the accuracy of risk level detection is 90%. Although the results were encouraging, the dataset was not annotated by experts, which is a fundamental criterion in legal text analysis.

### 3.2.2 Deep Neural Networks

As introduced in 2.4, deep neural networks are artificial neural networks with multiple layers between the input and output layers. A neural network can model any function and has non-linear activation layers to model non-linear functions. Due to the broad use of deep learning for legal text analysis, in this part, a brief literature review is presented.

Leveraging Recurrent Neural Network (RNN), [64] extracts obligations and prohibitions from contracts. The goal of this study is to assist legal firms and legal departments to automatically identify sentences (or clauses) specifying obligations and prohibitions in order to monitor the compliance of each party. The gold standard was compiled from the main bodies (excluding introductions, covers, recitals) of 100 randomly selected English service agreements. The NLTK's splitter<sup>10</sup> was applied to the 100 document and 31 545 training, 8 036 development, and 5 563 test sentences/clauses were identified. Five law students were selected to manually annotate the sentences/clauses with the five pre-defined classes: *Obligation*, *Prohibition*, *Obligation List Intro*, *Obligation List Item*, and *Prohibition List Item*. The results show that the best performance is achieved with a hierarchical BILSTM classifier, which produces an embedding vector for each sentence and then predicts a class for the sentence embeddings. Despite their promising results, the major limitation of this study is having only one expert opinion per class, which makes their dataset unreliable.

Neill et. al [65] employed Convolutional and Recurrent Neural Network to classify deontic modalities in regulatory documents. The annotations were carried out by Subject Matter Experts (SMEs) using the General Architecture for Text Engineering [66]. The final training set consists of 1 297 SME annotated sentences, including 596 obligations, 94 prohibitions, and 607 permissions. Furthermore, the test set consists of held-out documents from sub-domains of the financial regulations (e.g., Anti-Money

<sup>10</sup><http://www.nltk.org/>

Laundering, EU Markets in Financial Instruments Directive, Consolidated Accounts and Markets in Financial Instruments Regulation, etc.) and include 312 Obligations, 248 Permissions, and 62 Prohibitions. According to the paper, the inter-annotator agreement for two SMEs is equal to 0.74, with only a few disagreements. The results demonstrate that the NN model, which incorporates domain-specific legal distributional semantic model (DSM) representations with a general DSM representation (Google News), achieved the best performance. Although the conducted research in this study inspired us to combine domain-specific embeddings with a general one, the presented dataset was not useful for the contractual agreements domain.

Leveraging OPP-115 and deep learning, *Polisis* extracts segments from privacy policies and presents them to users in a visualized format [67]. According to the paper, the union-based gold standard was used for experiments, e.g., all experts' annotations were included in the gold standard (as opposed to the majority votes). Out of 115 privacy policies, 65 were considered for training, and 50 policies were kept for the test set. The authors claim that a successful multi-label classifier should not only predict the presence of a label but also its absence<sup>11</sup>. They report only macro-averages and further compute the average of F1 and F1-absence and yield 81% average on the test set. Despite the encouraging work done in *Polisis*, we believe that the paper lacks two fundamental elements: there is no validation set involved in the training phase; and there is no information on micro-averages.

It is worth mentioning that, regarding privacy policy classification, none of the above studies provided their dataset splits, and therefore there is no standardized benchmark for privacy policy classification. As a result, in Section 5, first, we show how we successfully reproduce *Polisis* results (though with different data splits) and further present two transformer models that significantly outperform *Polisis*.

### 3.3 Toward Violation Detection in Enterprise Policies

Policies are rules that are made by organizations to achieve their goals. In contrast, a regulation has the power of law and is a restriction that is imposed by authorities and leads to penalties for non-compliance. Since every policy and agreement should comply with the applicable law, in this Section, we study those efforts that attempt to detect inconsistency and violations in the human-readable policies.

In a recent study, Linden et al. conducted an experiment to analyze the privacy policy landscape after the GDPR enforcement [21]. The privacy policies of 3000 websites were fetched before and after the GDPR and were studied according to five criteria: presentation, readability, coverage, ambiguity and compliance. The analysis proves for positive changes triggered by the GDPR. The ambiguity level was improved in over 20.5% of the policies by using more precise phrases to describe the data collected and shared along with the purposes. In addition, an average of 15.2% of the policies improved their content across eight compliance metrics, e.g., they cover more data practices, particularly around data retention, user access rights, and specific audiences. However, the analysis reveals that there is a large gap between the current version of privacy policies and the ultimate goals of the GDPR. Despite the GDPR's emphasis on the right of the users to be informed, many of its requirements are still missing.

#### 3.3.1 Semantic Similarity Based Approaches

In order to compare the human-readable agreements with the regulatory documents, one established approach is to use semantic text matching. In essence, the objective of semantic text matching is close to anti-plagiarism efforts. Intelligent anti-plagiarism is an established field of study that exploits semantic

---

<sup>11</sup>They also claim that a model that predicts that all labels are present would have 100% precision and recall, which is obviously wrong.

similarity (more specifically, semantic text matching) to find potential violations. An extensive study of state-of-the-art has been conducted in [68]. 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 methods 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.

Sunkle et al. proposed an approach that uses Semantics of Business Vocabularies and Rules (SBVR) along with similarity measures to map the concepts of regulations to models of operational specifics of enterprises [69]. In addition to SBVR, the approach uses a terminological dictionary that includes vocabularies of operational specifics. Both vocabularies (SBVR) and the enterprise’s internal dictionary are converted to an intermediate formal representation, and then the state-of-the-art semantic similarity measure is used to compute the similarity and complete the mapping.

In his master thesis, Erl developed a recommender system to support regulatory experts at creating explicit references between semantically related controls from company policies and regulatory documents [70]. In order to compile a gold standard, the key statements from various regulatory documents of different topics (data privacy, fraud prevention, etc.) were manually extracted by the experts. The collection contains 879 controls statements that store explicit references to related controls from both regulatory documents and the company policies. Three different similarity algorithms were used: the classical TF-IDF, Word2Vec, and Doc2Vec, among which, Doc2Vec achieves the best results.

The limitation of the investigated studies is that, either the mapping process requires a remarkable manual effort by the experts or, in case of an existing reliable dataset, they are not publicly available. Thus, we set out to devise a new semi-automatic approach for mapping a contractual agreement to the legislation.

### 3.3.2 Machine Learning Based Approaches

In the last part of our literature review, we investigate the efforts that exploit machine learning to study regulatory compliance. CLAUDETTE is a Web server for the automatic detection of potentially unfair clauses in Terms of Service documents [71]. It employs machine learning and was trained on a corpus of 50 terms of service agreements. The dataset was annotated by lawyers and contains potentially unfair clauses. According to the paper, CLAUDETTE is able to achieve an average accuracy of around 80% in identifying potentially unfair clauses.

Founded on CLAUDETTE dataset, Lagioia et al. presented a methodology for detecting and explaining unfairness in consumer contracts using deep learning [72]. The dataset used for training and testing consists of 100 Terms of Services (ToS) of online platforms. The terms of services were analyzed and tagged in XML, based on eight pre-defined categories of clauses. The dataset contains 21 063 sentences, 674 of which include potentially or clearly unfair clauses. The paper specifically focuses on the *limitation of liability* clause and therefore faces a binary classification problem. The result shows that Memory Networks yield the best F-measure.

Last but not least, CLAUDETTE was extended to automate the legal evaluation of privacy policies using machine learning [73]. The evaluated corpus contains 3 658 sentences, where 401 sentences (11%) is marked as containing unclear language, and 1 240 sentences (33.9%) are tagged as potentially unlawful clause (i.e., a “problematic processing” clause, or an “insufficient information” clause). The experiment indicates that none of the analyzed privacy policies meet the requirements of the GDPR. The authors concluded that if a sufficiently large dataset is created, the vision of automatic compliance checking of privacy policies can be largely addressed.

# Semantic Interpretation of Contractual Agreements using Ontologies

---

In the previous chapters, we posed research problems, challenges, and questions. We also illustrated that since contractual agreements are a kind of legal text, they are written in a structured, although complicated terminology which make them an ideal use-case for Ontology-Based Information Extraction (OBIE). This chapter addresses a part of research question 1, where the focus is rule-based methods (vs. statistical techniques):

### Research Question 1 (RQ1)

Are text mining techniques able to extract valuable information from contractual agreements?

In addition, we designed quantitative and qualitative experiments to address the second research question in this Chapter:

### Research Question 2 (RQ2)

Does ontology-based information extraction help end-users to spend less time to understand contractual agreements?

Our running use-case for this chapter will be based on End-User License Agreements or EULAs. We present *EULAide*, which extracts permissions, prohibitions and duties from EULAs, clusters the similar-extracted excerpts based on semantic similarity and provides a structured summary to the users.

The contributions of this chapter towards the stated research questions in general are:

- A pipeline for EULA information extraction comprising linguistic pre-processing, the inclusion of an ontology-based gazetteer, and a large number of custom extraction rules.
- Implementation of a license analysis framework, which comprises a web API and UI.
- A comprehensive evaluation of our approach benchmarking it against human judgment.



Chapter 4 is based on the following publications [74, 75]. The Chapter is structured as follows: in Section 4.1 we explain our approach and its implementation in the *EULAide* Service; Section 4.2 evaluates the OBIE approach and the clustering alternatives as well as the usability of *EULAide* service; and finally Section 4.3 examines the research questions and provides concluding remarks.

## 4.1 Ontology-Based Information Extraction from License Agreements

We introduce a technique that exploits knowledge encoded in an ontology for annotating, extracting, and classifying EULA content into pre-defined categories, leveraging ontology-based information extraction. OBIE guides Information Extraction (IE) pipelines to process unstructured or semi-structured natural language text by exploiting ontologies to extract pre-defined structured information and annotating the text using ontology terms. OBIE was favored for several reasons. Primarily, the reliance on a vocabulary engineered by domain experts grounds our work in existing standards and broadens its application. Secondly, as a form of a legal document, license agreements tend to have clear structures and terminologies, sometimes even containing identical clauses and phrases. This facilitates the ‘mapping’ of natural-language text to machine-readable conceptualizations in the ontology for our use-case.

Following a survey of existing vocabularies, presented in Table 4.1, we identified the *Open Digital Rights Language* (ODRL) ontology (introduced in Section 2.3.1) as the most appropriate basis, based on its maturity and comprehensiveness. Despite being created specifically for digital content, the ODRL is broad enough to be used for different types of resources (such as linked open data, digital works, online services, etc.). As represented in the table, the domain coverage ranges from very specific (MPEG-21, e-commerce applications) to very broad (digital rights, open data, linked data). ODRL stands out not only in terms of being the most current (most recently updated vocabulary) but also in terms of comprehensiveness (ranking 2nd from the domain-independent vocabularies). ODRL is also a W3C recommendation and has demonstrated the highest community endorsement. Examples included the RDF licenses database<sup>1</sup>, which is a first attempt at developing a knowledge base of licenses, and which combines the Creative Commons Rights Expression Language (CC REL) and ODRL to express EULAs as RDF. Furthermore, extra efforts have identified ODRL as the best candidate for defining policies in linked data licenses [76]. In [77] a formal conceptual model based on CC REL, XrML and ODRL was integrated into a platform. The license picker ontology also exploits ODRL [78].

Similarly, out of several available Natural Language Processing (NLP) tools supporting OBIE methods, the GATE framework coupled with its ANNIE IE system [66] and its support for customized Java Annotation Pattern Engine (JAPE) grammar rules [84] was identified as the most appropriate for the following reasons. First, OBIE is supported even at the level of hand-coded grammar rules (via JAPE). Second, it offers one of the easiest methods for embedding a GATE pipeline in Java. Third, it has strong evaluation tools for NLP, including inter-annotator agreement and F-measure calculation based on a gold standard, thus relieving the developer from combining different software solutions for one single task. In the next subsection we describe the architecture and implementation of the *EULAide* system, which builds on the GATE framework.

### 4.1.1 Architecture

Figure 4.1 shows the architecture of the framework. There are two main components in the back-end: the GATE OBIE pipeline and the clustering module. Each of the following subsections will explain how

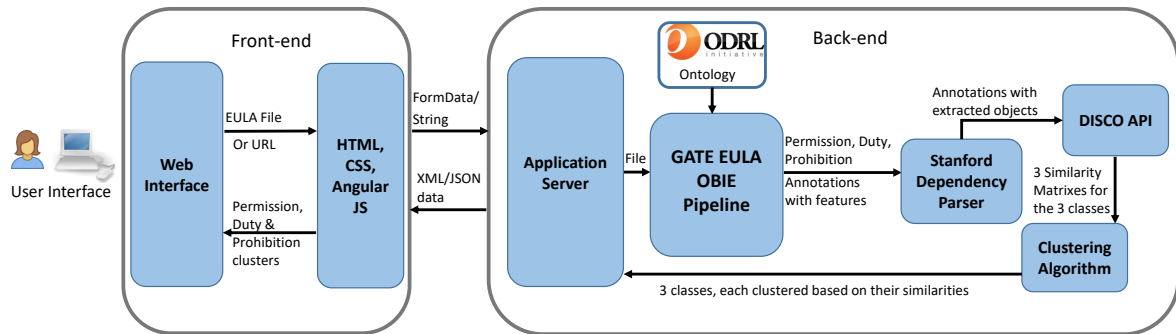
---

<sup>1</sup><http://datahub.io/dataset/rdflicense>



Table 4.1: Vocabularies and ontologies for EULAs.

Name	Domain Coverage	# Classes & Instances	# Properties	Last Release
<i>CC REL</i> [28]	linked data	28	42	2013/11
<i>ODRL</i> [29]	open digital content	85	56	2018/02
<i>LDR</i> [79]	linked data resources	77	21	2014/09
<i>LALOD</i> [80]	web of data	16	5	2013/05
<i>ODRS</i> [81]	open data	2	15	2013/07
<i>MPEG-21</i> [30]	contains the terms as standardized in ISO/IEC 21000-6	2000 standardized terms having the characteristics of a structured ontology		2005/07
<i>Copyright Onto</i> [82]	digital rights management	99	42	2019/09
<i>IPROnto</i> [83]	intellectual property rights with a focus on e-commerce applications	113	54	2003/12

Figure 4.1: Architecture of the *EULAide* system.

each contribution fits within this framework, i.e., i) GATE pipeline for OBIE, ii) word space creation for the proposed clustering method, and iii) the *EULAide* service itself. The latter fits within the front-end, the former two within the back-end shown.

#### 4.1.2 GATE OBIE Pipeline

GATE provides three types of resources: *Language Resources* (LRs) which collectively refers to data; *Processing Resources* (PRs) which are used to refer to algorithms; and *Visualization Resources* (VRs) which represent visualization and editing components. Figure 4.2 illustrates all the PRs which are specifically tailored for EULA processing. The inputs for this pipeline is a EULA in natural language text and the ODRL ontology. The pipeline consists of (1) a linguistic pre-processing stage, (2) an ontology-based Gazetteer, (3) the primary OBIE transducer, and finally (4) the feature extractor for the clustering. The *Linguistic pre-processor* consists of the following PRs:

- *Tokeniser*: adds two annotation sets, e.g., *Token* and *spaceToken*.

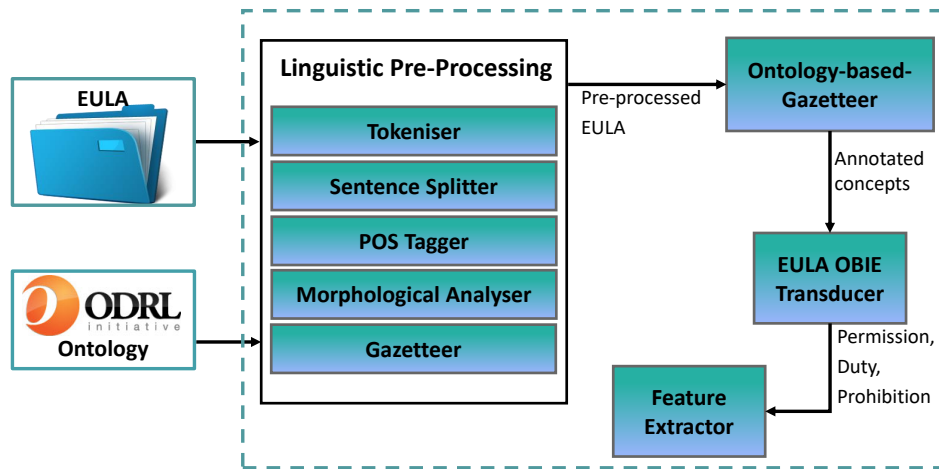


Figure 4.2: GATE EULA OBIE Pipeline.

- *Sentence Splitter*: splits the sentences and creates the *Sentence* annotation set.
- *POS Tagger*: applies Part-of-Speech tagging and adds a feature called *category* to each *Token* annotation.
- *Morphological Analyser*: inserts a new feature to each *Token*, called *root*. Later the *ontology-based Gazetteer* annotates the concepts based on the *root* feature.
- *Gazetteer*: reuses existing relevant lists (e.g., countries) but adds additional lists covering terms that carry important information in license agreements, such as file formats and different synonyms for the term ‘license’ and ‘asset’.

The next major pipeline component is the *Flexible Ontology-based Gazetteer*, which takes the ODRL ontology as input and creates a new annotation set called *Lookup* containing matches to ODRL instances and concepts. Basically, this PR matches any text features to the ODRL element labels, flexibly allowing for various inflections. The result is a list of semantic annotations pairing bits of text to the matching ODRL elements.

The most crucial PR is the customized *EULA OBIE Transducer*, which considers all the previous annotation sets as inputs and matches pre-defined annotation patterns to the final annotation sets: (Permission, Prohibition, Duty). This PR is described in detail below.

### EULA OBIE Transducer

The transducer executes in 10 phases and builds on all outputs from the previous stages to create the annotations. We have implemented 15 grammar rules to generate the final *Permission*, *Prohibition* and *Duty* annotation sets. The definition of the JAPE rules is heavily based on ODRL community specification documentation, where each class and subclass is explained in detail. For instance, according to the vocabulary documentation, *include* is an instance of *Action* class and means: “*The Assigner requires that the Assignee(s) include(s) other related assets in the Asset.*”. Therefore, the presence of ‘include’ in a sentence can suggest the presence of a *Duty* in a EULA.

In order to extract `Actions` more precisely, we added some rules in GATE morphological PR. This resource specifies the root of each token, and in most cases, the stems of nouns are identified almost as the original noun itself, e.g., the lemmas for ‘distributions’, ‘attribution’ or ‘attachment’ are ‘distribution’, ‘attribution’ and ‘attachment’. In this regard, the OBIE pipeline can not relate these words to the ontology concepts, because the *Ontology-based Gazetteer* annotates the text based on the root of each token. However, after customizing the morphological analyzer, the accuracy of stem identification has improved significantly. For 20 EULAs, the number of annotated concepts by the ontology-based Gazetteer has increased from 9 630 to 9 927. As a result, the *Ontology-based Gazetteer* is now for example able to extract the following `Action` in: “Activities other than **distribution** and/or **modification** of the Work are not covered by this license..”. In the remaining section, we explain the main phases of OBIE Transducer in more detail:

**annotateClasses** This phase separates the *Lookup* annotation set, which contains all the ontology-derived annotations. In order to extract valuable information from EULAs, we have focused on the `Rule` class defined by ODRL, since it is an “abstract common ancestor to `Permission`, `Prohibition` and `Duty` classes”. Some properties of `Rule` include `action` (the operation relating to the asset) and `constraint` (constraints which affect the validity of actions). Since the three relevant subclasses of `Rule` inherit these properties, the ODRL ontology satisfies our needs. According to ODRL community group explanations<sup>2</sup>, we have differentiated two main categories for actions: *DutyAction* and *Permission-ProhibitAction*. Although some actions are present in both annotations (like `delete`), this separation is a vital step for the next phases. Apart from actions, essential words that carry significant information for rights detection are also determined. For instance, *must* and *should* are labeled with *DutyWords*; *may*, *can*, *grant*, *permit*, etc. are labeled with *PermissionWords* and similarly *may not*, *can not*, *not allowed*, *prohibited*, etc. are labeled with *ProhibitionWords*.

**extractPermissions** Since there may be different structures of sentences in a license, we implemented four rules for the extraction of permissions. For instance the sentence “[You] [may] [copy, share and reproduce] [the product]” will fire the following grammar rule (+ means one or more occurrences): `[Subj] [permWords] [permAction]+ [Asset]`. On the other hand, the sentence “[This license] [grants] [you] [to copy, share and reproduce] [the product]” will fire another rule: `[License] [permWords] [Object] [permAction]+ [Asset]`.

It should be clarified that some annotation sets such as *License* and *Asset* are detected by our own-defined *gazetteers*. Table 4.2 shows the different steps towards extracting of the above `Permission`. After the pre-processing phase, first the *ANNIE gazetteer* generates two annotation sets: *License* and *Asset*. Second, the *ontology-based gazetteer* annotates the concepts based on the ontology with *Lookup* label. Then the first phase of *EULA OBIE transducer* is executed and the *PermWords* and *PermAction* annotation sets are created. Finally, the *PermissionI* rule from the second phase fires and annotates the whole sentence as a `Permission`.

**extractProhibitions** This phase is very similar to the previous one, except that in the grammar rules the *PermissionWords* are replaced with *ProhibitionWords*. Therefore the sentence “[You] [may not] [copy, share and reproduce] [the product]” will be annotated as a `Prohibition`.

**extractDuties** To extract `duties`, there are more diverse structures. Hence we implemented five different rules, one of which is the following: `[Subj] [DutyWords] [DutyAction] [obj] [Asset]`

<sup>2</sup><https://www.w3.org/TR/odrl/>

Table 4.2: Example of a Permission as extracted by *EULAide*.

<i>ANNIE Gazetteer</i>	<b>This license</b> License	grants	you	to copy, share and reproduce	<b>the product</b> Asset
<i>Ontology based Gazetteer</i>	This license	grants	you	<b>to copy, share and reproduce</b> Lookups	the product
<i>Annotate Classes Phase</i>	This license	<b>grants</b> Perm Words	you	<b>to copy, share and reproduce</b> Perm Actions	the product
<i>Extract Permissions</i>	<b>This license</b> License	<b>grants</b> Perm Words	<b>you</b> Obj	<b>to copy, share and reproduce</b> (Perm Actions)+	<b>the product</b> Asset

This rule fires when processing the sentence: “[You] [must] [include] [a copy of this License] [with your product]”.

**clean** In this phase, all intermediate annotation sets are deleted from the output and only the three final annotation sets are retained: `Permission`, `Prohibition`, `Duty`.

### Feature Extraction for Clustering

In an empirical study of 20 common licenses, we observed that many policy excerpts returned for each class could be thematically grouped into clusters. As an example, table 4.3 shows three segments which have been extracted as `permissions` for the Apache License. The colored words have the same or very similar meaning and can, therefore, be grouped together. It should be clarified here that our approach does not intend to remove any extracted segment from similar ones, since this may lead to losing vital information in EULAs. Instead, our goal is to provide a summary for each cluster. If the end-user is concerned about a specific policy, they can browse the list of items in each cluster and see the details.

The last component in Figure 4.2 is the feature extractor. We have taken `permissions`, `duties` and `prohibitions` excerpts as the input and extracted key features from them in order to perform clustering. The extracted features carry crucial information in EULAs and play a fundamental role in clustering similar segments. Features include:

- the sequence of `action` for each segment, e.g., ‘copy, reproduce’, ‘share’, ‘remove’, etc.;
- the `condition` on which a specific `action` is granted or forbidden or obliged; and
- the `typeOfPolicy` which can be a ‘copyright’ or ‘patent’ or ‘intellectual property right’.

Table 4.4 shows two examples of expected results of the feature extraction phase. Once the three annotation type classes are built with the respective features, they are passed to a semantic similarity measurement component. This component builds a symmetric matrix for each class and passes them to the clustering algorithm. Finally, the clustering component groups the segments based on their similarities and the clustered `permissions`, `prohibitions` and `duties` are shown to the end-user. The next section provides more details regarding the similarity computation and its usage for text clustering.

Table 4.3: Example of annotated permissions in Apache.

You may <b>reproduce</b> , prepare <b>Derivative Works</b> of, publicly display, publicly perform, <b>sublicense</b> , and <b>distribute</b> the <b>Work and such Derivative Works</b> in Source or Object form.
You may <b>reproduce</b> and <b>distribute</b> copies of the <b>Work or Derivative Works</b> thereof in any medium, with or without modifications, and in Source or Object form.
You may <b>add Your own attribution notices</b> within <b>Derivative Works</b> that You <b>distribute</b> , alongside or as an addendum to the NOTICE text from the Work.

Table 4.4: Example of features extraction.

If you join a Dropbox for Business account,	you must	use	it in compliance with your employer's terms and policies
<b>condition</b>		<b>action</b>	
each Contributor grants to You a	patent	license	to make, use, sell, import, and transfer (the Work)
	<b>type of policy</b>		<b>action</b>

### 4.1.3 Word Space Creation and Semantic Clustering

The clustering method exploits semantic similarity between word and text for grouping similar excerpts. Semantic text similarity uses word embeddings and distributional approaches. Distributional semantic approaches benefit from the observation that words in similar contexts tend to have similar meanings. Consequently, no labeled data is necessary for learning the embedding vectors, but only great amounts of correct text. Once the important segments and their features are extracted, a semantic similarity framework based on word embeddings is used to compute the similarity between different features of each class (Permission, Prohibition and Duty). Afterwards, by summation of features similarities, the final similarity score for each pair is calculated, and a symmetric similarity matrix is built for each class. Finally, the most similar segments in each class are grouped together with hierarchical agglomerative clustering (HAC) algorithm, and the procedure goes on until a certain threshold is reached.

DISCO (extracting DIstributionally related words using CO-occurrences) is an open-source Java application that retrieves the semantic similarity between short texts and phrases [85]. Besides, it allows users to build their own word embedding database from a text corpus. Over the years, DISCO has received a high community endorsement [86–89]. Thus, we combined the OBIE method with the DISCO clustering method for EULA summarisation.

In order to create a domain-specific word embedding space for our approach, we used a dataset comprising 1000 EULAs [90]. The dataset is passed to a method that generates a lemmatized text file consisting of three columns: token, a part-of-speech tag, and the base form (lemma). We executed the DISCO builder with the default configuration on the lemmatized file. The builder's output contains word vectors for each token. Finally, DISCO takes word space and two short texts as input and generates a real value between one and zero, indicating the semantic similarity.

For computing the similarity between two short texts, DISCO uses the following formula:

$$Sim(T_1, T_2) = \frac{directedSim(T_1, T_2) + directedSim(T_2, T_1)}{2} \quad (4.1)$$

Assuming  $T_1$  and  $T_2$  consist of  $n$  word corresponding to  $w_{11}, \dots, w_{n1}$  and  $w_{12}, \dots, w_{n2}$ , the directed similarity is calculated as shown in equation 4.2. The function  $weight(word)$  returns a real value (between

---

**Algorithm 1** Sketch of semantic clustering algorithm.

---

**Require:** permissions, prohibitions, duties with features

- 1: **for** the three classes **do**
- 2:     **for** all segment pairs in each class **do**
- 3:          $A \leftarrow$  similarity between actions ▷
- 4:          $B \leftarrow$  similarity between conditions ▷
- 5:          $C \leftarrow$  similarity between policy types ▷
- 6:          $D \leftarrow$  similarity between the remainders of segments ▷
- 7:          $finalSim \leftarrow A + B + C + D$  ▷
- 8:         add finalSim to the corresponding matrix cell ▷
- 9:     do HAC clustering for the matrix with a threshold

**Ensure:** clustered permissions, prohibitions & duties

---

0 and 1) for an input word. The weighting algorithm is based on frequency of the word in the corpus.

$$directedSim(T_1, T_2) = \sum_{i=1}^n weight(w_{i1}) * \max_{1 \leq j \leq n} [WordSim(w_{i1}, w_{j2})] \quad (4.2)$$

Algorithm 1 shows a sketch of our clustering approach. The semantic similarity is computed for all features of the extracted segments, and the final similarity score is calculated by summing all four values. The rationale behind the summation operation is to be in line with the formula in equation 4.2. Once we obtained the similarity matrix for each class, a clustering algorithm is required to group similar segments. Agglomerative hierarchical clustering (HAC) is an established, well-known technique that has been shown to be a successful method for text and document clustering [91]. Furthermore, among different HAC methods, the average linkage has been proved to be the most suitable one for text categorization [91, 92]. Once the proper clustering technique is identified, we can pass similarity matrices to the clustering component. The HAC process continues until it reaches a pre-defined threshold.

#### 4.1.4 EULAide Framework and Web Service

We developed a comprehensive license analysis framework *EULAide*, which also provides a comprehensive Web service interface. As shown in the architecture, the client-side Web interface implemented in Javascript using the AngularJS framework sends a request to the back-end service and receives a JSON or XML object if the request is valid. In the back-end, several open-source APIs are orchestrated to provide an accurate result.

As demonstrated in Figure 4.1, the architecture contains a dependency parser that is responsible for resolving the main object of each excerpt. In such a manner, we are able to generate a short and simple summary for each cluster by concatenating the ‘action’ and ‘object’ of all segments in one cluster. It is worth mentioning that integrating GATE with the Stanford NLP was quite a challenge since both have defined quite different structures for annotations and related concepts in their APIs.

Figure 4.3 shows an example of *EULAide* output, applied to the Google terms of service<sup>3</sup>. The number of extracted excerpts by the OBIE pipeline is 14, whereas the number of clusters has reduced to 9. The head of each accordion (block) contains the concatenation of ‘action’ and ‘object’ (extracted by Stanford dependency parser) of all members in the cluster, which can be replaced with a more

---

<sup>3</sup><https://www.google.com/policies/terms/>

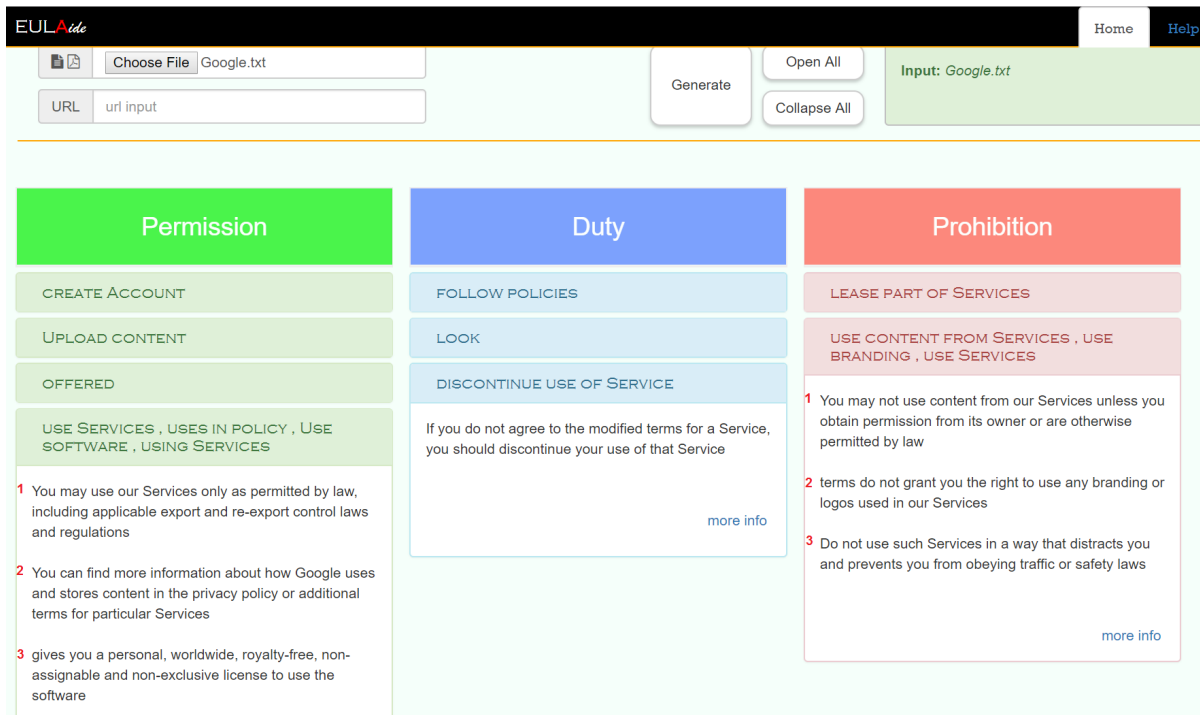


Figure 4.3: *EULAide* platform Web interface showing the permission, duty and prohibition clusters for a user provided EULA.

proper summarization algorithm in future. As represented in the figure, the numbered excerpts are grouped together. In addition, there is a tool-tip "more info" at the bottom of each block, which the user can hover over and see the complete paragraph regarding that policy. The user can also see all the details by clicking on the "Open All" button or choose to expand a specific accordion by clicking on the corresponding header (e.g., summary). Since *EULAide* is implemented using a two-layer architecture, it is platform-independent, e.g., any client such as mobile apps can easily communicate to our server.

Figure 4.4 shows a sample mockup where a native iOS app communicates to the *EULAide* server by sending a text file including terms and conditions of iTunes store and retrieves a JSON response containing all clustered duties, permissions and prohibitions along side with the cluster summaries. In [93], an approach is presented founded on *EULAide*, with the aim of optimizing user's experience of EULA interpretation on mobile devices. The author developed two native mobile apps for iOS and Android using React Native [94]. The app has two main screens where the first screen allows the to enter the URL of the EULA, and the second screen presents the summary, which is retrieved by an HTTP request to the server.

## 4.2 Experimental Study

This experimental study aims to evaluate to what extent *EULAide* is able to extract valuable information from license agreements and whether the extracted segments assist end-users in familiarizing themselves with the terms and conditions of an arbitrary service.



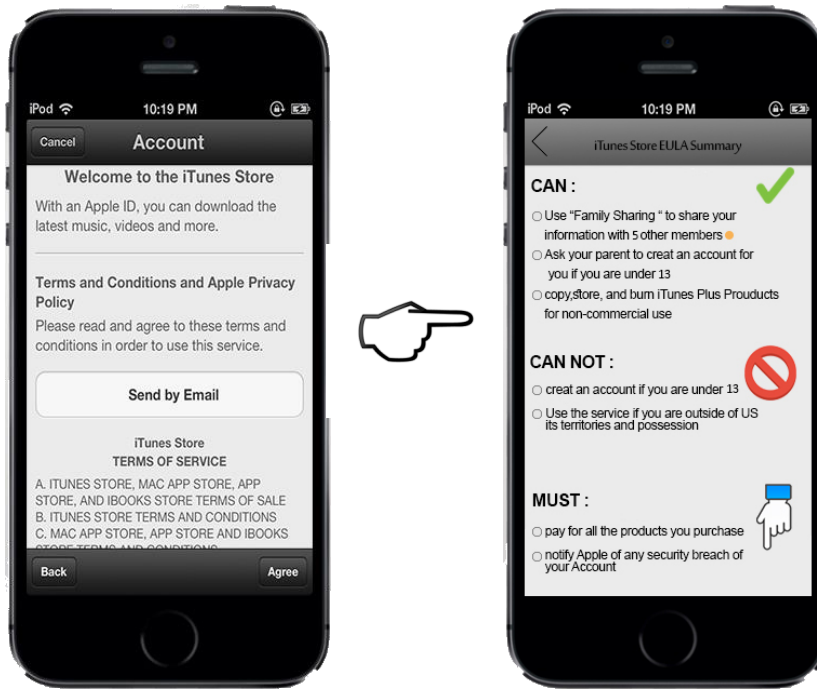


Figure 4.4: A sample mockup of a mobile app communicating to the *EULAide* Web service.

#### 4.2.1 OBIE Pipeline Evaluation

In this section, we compare the OBIE pipeline performance against our compiled gold standard. Currently, our method does not require any training set, and the information extraction relies solely on the described JAPE rules and gazetteer entries derived from the ontology. However, a solid test set was required in order to:

- manually extract relevant grammatical and lexical patterns and translate them into JAPE rules;
- use the test set as a base for an inter-annotator agreement test to identify realistic upper-bounds of successful automatic extraction;
- evaluate the OBIE approach by comparing the pipeline execution F-score to the level of inter-annotator agreement achieved.

In the next sections, we explain how we set about to achieve a gold standard following an inter-annotator agreement experiment, and discuss the evaluation setup and results of the OBIE approach.

#### Gold Standard Creation

Although some EULA datasets are available, neither of them is annotated at the required level of granularity. For example, some RDF expressions of EULAs are freely available<sup>4</sup>. However, they were not useful for our experiments for two reasons. Primarily, a comprehensive dataset should contain two kinds of licenses: license templates (e.g., the core definition of the GNU Public License) and actual instances (e.g., an Apple Inc. EULA). Unfortunately, the available dataset includes only the core definitions.

<sup>4</sup><http://datahub.io/dataset/rdflicense>



Table 4.5: Specification of end-user license agreements.

Row	Name	Word Count	Character Count (excl. whitespace)
1	Apache	1 581	8 652
2	Apple Website	3 328	19 831
3	bitTorrent	4 383	23 215
4	Dropbox	1 938	9 859
5	Eclipse	1 701	9 534
6	EUP License	2 087	10 709
7	Facebook	4 404	22 435
8	GNU	5 614	28 386
9	Google	1 869	9 501
10	iTunes Boss	965	5 272
11	Jetbrains	1 833	10 230
12	LaTeX	3 011	15 377
13	Minecraft	1 962	8 489
14	Mozilla	2 821	14 829
15	Netflix	5 134	27 097
16	Python	1 440	8 034
17	Red hat	3 303	17 085
18	Skype	4 365	22 571
19	SoundCloud	8 927	46 779
20	Sun	3 450	18 406

Secondly, there are no structural links between the annotations and the text, and therefore we could not use it for our evaluation purposes. The only annotated corpus that was suitable for our objectives from a requirement point-of-view was the NLL2RDF project dataset<sup>5</sup>. However, despite containing 37 annotated licenses, the frequencies of annotations are too low, and the annotations were deemed incomplete if not unreliable. For instance, no `Prohibition` or `Duty` annotations, and only four `Permission` annotations were observed for the Mozilla license. In contrast, after a careful manual annotation we identified and verified 16 `duties`, 12 `permissions` and 3 `prohibitions`.

In order to compile our own gold standard, we collected 20 popular EULAs in their original text<sup>6</sup>. Table 4.5 shows the details of the input set, including the average word and character count. When choosing the licenses, we intentionally tried to cover varying ones both in terms of structure and content. For instance, since Mozilla and Netscape are offered by the same organization, their EULAs carry a lot of identical phrases. Furthermore, we selected licenses of varying lengths, purposely avoiding ones that are too short. The average word count of a license within the corpus is 3 206, and the average character count without space is 16 815.

To prepare the gold standard, two annotators familiar with EULA-like text annotated the corpus independently following an introduction to the relevant ODRL concepts. Using the dedicated GATE plugin, we computed the Inter-Annotator Agreement (IAA) score for the two annotation sets. The plugin offers two types of IAA measurement: F-measure and agreement based on the kappa statistic. The

<sup>5</sup><http://www.airpedia.org/nll2rdf/dataset-licenses/>

<sup>6</sup>Although a few chosen EULAs are terms of services, they were selected for our experiments, since they contain digital rights and conditions.

Table 4.6: Lenient IAA for two annotators.

	Precision	Recall	F-measure
<b>Permission</b>	0.94	0.90	0.92
<b>Prohibition</b>	0.79	0.94	0.86
<b>Duty</b>	0.86	0.96	0.91
<b>Summary</b>	0.87	0.93	0.9

latter has been criticized and has some well-known limitations. Kappa is suitable when annotators have the same number of instances but with different class labels. It is not recommended for text mark-up tasks, such as named entity recognition and information extraction [95]. When the annotators themselves determine which text spans they can annotate, the F-measure should be used. The F-measure has been less controversial and is also indicated as the most appropriate IAA measure in the GATE manual itself, given the nature of our annotation task [66].

GATE offers three ways to measure IAA: the *Strict* measure only takes the annotations that have precisely the same span in the text and considers all partially correct annotations as incorrect; the *Lenient* measure considers all partially correct annotations as correct; and the *Average* measure allocates a half weight to partially correct annotations (i.e., it takes the average of strict and lenient). Since our goal is to extract blocks of text representing pre-defined concepts, we do not require fine-grained results. Therefore, the lenient measure was deemed sufficient since we simply want to guide human readers to which parts of the text contain `permissions`, `duties`, and `prohibitions`.

An initial lenient IAA resulted in an acceptable F-score of 77%. Considering the complexity of EULA texts, the initial IAA is well within the reasonable performance range described in related literature for IE tasks of a similar complexity [96]. The annotators were then invited to discuss their disagreements with the aim of conflict resolution. The results of this discussion also contributed to additional improvements to our customized gazetteer and JAPE grammars. After this consultation phase, the IAA increased to a very satisfactory 90%. Table 4.6 shows the lenient IAA results. In order to produce a final gold standard for evaluating the actual *EULAide* pipeline, we removed the 10% disagreements and only retained the agreed-on annotations. The final gold standard contains 193 `permissions`, 185 `prohibitions` and 168 `duties`.

## Evaluation & Discussion

In this section, the conducted experiments are presented. We have utilized the Corpus Quality Assurance tool in GATE. The tool calculates precision, recall, and F-score between two annotation sets in a corpus. Similar to IAA, lenient F-measure was selected for the comparison of *EULAide* with the gold standard. Furthermore, among different types of F-scores, we decided to rely on F2-score, since in the legal document’s domain and EULA as a specific type of legal texts, it is hazardous not to detect some `prohibitions` and `duties`. For instance, if there is a license agreement in a hospital, it is crucial to identify all types of `prohibitions` and `duties`, and missing important policies may cause adverse effects. However, in order to provide a complete overview of *EULAide* performance, all three types of F-scores are reported here.

Tables 4.7 shows the evaluation results. As represented in the table, the precision for the `Permission` class is 74%. One of the existing false positives from *Eclipse*<sup>7</sup> license is the following sentence: “*Everyone is permitted to copy and distribute copies of this Agreement*”. Although this excerpt seems to

<sup>7</sup><https://www.eclipse.org/org/documents/epl-v10.php>

Table 4.7: Evaluation of OBIE pipeline.

	<b>Precision</b>	<b>Recall</b>	<b>F0.5</b>	<b>F1</b>	<b>F2</b>
<b>Permission</b>	0.74	0.75	0.74	0.74	0.75
<b>Prohibition</b>	0.89	0.63	0.82	0.74	0.66
<b>Duty</b>	0.66	0.67	0.67	0.67	0.67
<b>Overall</b>	0.75	0.68	0.74	0.72	0.7

contain a `Permission`, the `Permission` is not related to the asset (which in this case is the software) and is a grant about copying/distributing the natural language license<sup>8</sup>. Furthermore, according to the recall value in the table (75%), *EULAide* has missed some permissions due to the incomplete set of instances in the ontology. For example, this segment from EUPL<sup>9</sup> is annotated as a `Permission` in the gold standard: “*The Licensor hereby grants You to sublicense rights in the Work or copies thereof*”. In this case, ODRL does not contain `sublicense` instance, and therefore the ontology-aware Gazetteer misses this annotation. The problem arises from the fact that ontologies can not always cover all the requirements. Most ontologies are useful because they provide the basic modeling, but they need to be extended for some use-cases.

For the `Prohibition` annotations, we achieved a precision of 89% and a recall of 63%. Some of the incorrect prohibitions annotated by *EULAide* can be traced back to the annotators’ disagreements in the IAA experiment. As an example, one of the disagreements in the IAA was the following sentence: “*Do not use such Services in a way that distracts you and prevents you from obeying traffic or safety laws.*”. While one of the annotators believes that this is a soft `prohibition` for the user, the other considers this phrase as a warning to the licensee and believes that this sentence does not prohibit the users from accessing the service. However, since it was not a both-agreed annotation, it was removed from the gold standard and therefore is counted as one of the incorrect annotations generated by *EULAide*. In addition, to increase the recall for the `Prohibition`, we should consider the language variability in EULAs. In implementing the rules we have realized most `prohibitions` include the terms “*You must/should/may not*”. However, unsurprisingly some of them have more complex structures (e.g., “*No one other than Sun has the right to modify the terms applicable to Covered Code created under this License*”). Therefore, including different possible structures of a sentence in a rule would lead to a higher recall.

Finally, the precision for `Duty` annotations is 66%. While we had assumed that the terms “*it is your responsibility*” or “*you are responsible*” along with some other patterns would lead to a `Duty` in a EULA, the annotators did not mark all of them as `duties`. For instance, *EULAide* annotated the sentence “*Each Recipient is solely responsible for determining the appropriateness of using and distributing the Program*” as a `Duty`. On the contrary, the annotators believe that this is more like a warning to the user and does not mean an obligation the user should satisfy in order to receive a specific `Permission`. Last but not least, the recall value for `Duty` is 67% and shows that *EULAide* did not detect 33% of the `duties`. Similar to the `Prohibition`, several structures of a natural language make the IE process complex. For example, in most EULAs, the licensor usually uses the words “*You must/should/have to*” for defining a `Duty`, but on the other hand there are some with different forms

<sup>8</sup>The benefit of applying rule-based information extraction is that one can learn from wrong annotations and improve the implementation of grammar rules. However, changing the grammar rules based on the gold standard will lead to implementation bias. Therefore the rule enhancement can only be effective when there is a new and fresh gold standard against which the enhanced version can be tested. Since compiling another gold standard is a tedious and expensive task, we decided not to change the grammar rules.

<sup>9</sup><https://opensource.org/licenses/EUPL-1.2>

stating an obligation (e.g., “*The GNU GPL requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions*”). Catching the name of EULAs and passing them to the JAPE rules is one possible solution to detect such `duties`.

### 4.2.2 Evaluating the Clustering Approach

Over the years, many clustering methods have been proposed for the summarization of texts. Survey papers such as [91] indicate a large variety of methods can be pursued. Efforts investigating semantic similarity for summarization purposes have either relied on a corpus-based approach or have implemented a new algorithm to compute the semantic similarity between words and short texts. Kenter & Rijke investigated whether it is possible to compute the similarity between two short texts relying on just semantic features of the texts [97], using machine learning methods. Tsatsaronis et al, proposed a new approach for computing semantic relatedness between words based on a word thesaurus [98]. A framework presented in [99] proposes semantic role labeling for the summarization task. In a similar study, Aliguliyev [100] proposed a sentence-clustering approach for extractive document summarization, using an evolutionary algorithm and an appropriate ranking method.

Although examples like the above abound, the uniqueness of our target subject data precludes any method from being considered superior. Due to the nature of the target subject data, we have considered the semantics-based clustering of segments as a base for an intuitive Web user interface that supports human analysis of licenses. The evaluation carried out, investigates two hypotheses:

- to identify whether the proposed clustering method based on the OBIE-derived classes is generally useful (i.e., helps readers better digest and comprehend EULAs) and;
- to identify whether the devised feature-extraction method offers improved results;

### Experimental Setup

When available, the quality of clustering methods is best measured by comparing machine-generated clusters with a reliable gold standard. However, considering the complexity and broadness of EULAs, it is challenging to obtain or compile a suitable gold standard that is agreed upon and accepted by a majority. Therefore, as an alternative method, we designed an experiment that can compare clustering preferences between human evaluators (to approximate inter-annotator agreement), and compare them with those generated computationally. To determine the usefulness of feature extraction, we included a second machine-generated clustering for a secondary comparison.

The input data for the evaluation consists of a number of instances for the three EULA classes, `prohibitions`, `duties` and `permissions`, from four carefully-selected EULAs. The choice of the latter considered both a good balance between the three identified classes as well as sufficient brevity. EULAs that were very long were not suitable for this task since the human annotators’ clustering task increases significantly in terms of complexity. The OBIE pipeline generates the total number of class instances shown in table 4.8 for the four-selected EULAs. Here, one can note that the feature-based clustering method yields a marginally higher amount of clusters for two of the three classes (e.g., `Permission` and `Prohibition`). A higher amount of clusters can be interpreted as a more fine-grained result. However, this can only be confirmed by comparison of these results with those of our human evaluators.

To carry out human evaluation, five subjects were asked to cluster the OBIE-derived excerpts of `Permission`, `Prohibition` and `Duty` within the context of each selected EULA. Since the target users of *EUALAide* are regular people who are not expected to be particularly acquainted with legal text

Table 4.8: Total extracted instances &amp; machine-generated clusters with (Mf) and without (M) considering features.

	#Instances	#Clusters- M	#Clusters-Mf
<b>Permission</b>	30	18	20
<b>Duty</b>	27	24	20
<b>Prohibition</b>	40	32	35

and jargon, the five volunteers selected have different levels of higher education (under- and postgraduates) selected from a university campus. The only confirmed common interest between the individuals is an understanding of the need for EULA summarization methods, such as the ones we propose. At the same time, to ensure that the task is properly and equally understood by all evaluators, an introduction to the *EULAide* tool, its vision, goals, and the relevant concepts behind the input data was provided. However, the evaluators were not given instructions on how to cluster the results but were rather asked to devise their own clustering criteria as they best deemed fit. They were also explained that the excerpts were semi-automatically extracted and could, therefore, contain some errors. The EULAs they considered contained and average of 7.5 permissions, 10 prohibitions and 6.7 duties.

### Human-Machine Comparisons & Discussion

The experiments presented yielded two result sets: i) the two machine-generated clusters (with and without considering features) and ii) the five human-generated clusters. In order to compare and consider the two sets, we considered conventional methods for measuring clustering quality, i.e., the Rand index and F-measure [101]. Applying F-measure to clustering methods is similar to other information retrieval approaches, and considers (dis-)similarity of excerpts within clusters as a base for true/false positive/negatives. However, the F-measure disregards true negatives, or “the occurrence of dissimilar excerpts in different clusters”. Thus, it does not take the proportion of correct non-clustering of unrelated excerpts into account; which is also very important in measuring the success criteria for this task. Alternatively, the Rand index formula, shown in equation 4.3, is used to measure the percentage of accuracy, which is more appropriate for our evaluation because it also factors true negatives.

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \quad (4.3)$$

The interpretation of positive or negative decisions in clustering is drawn from decision series theory in arithmetic. Having  $n$  items, the space of all pairs of elements is computed by equation 4.4.

$$\binom{n}{2} = \frac{n * (n - 1)}{2} \quad (4.4)$$

In our results set, we do not have one ‘correct’ clustering, but rather five subjective variations from each evaluator. Therefore, we applied the Rand index to calculate the cross-accuracy of clusters, using each evaluators’ clusters in turns as the correct standard. We then simply considered the two machine-generated clusters as alternate clusters and extended the cross-accuracy computations to cover both result sets. These computations generate a (symmetric) matrix of results, separated by class (permissions, duties and prohibitions), as shown in table 4.9. The five human evaluators are enumerated as h1-h5 and the two machine-generated clusters as M (without feature inclusion, i.e., passing the entire excerpts to the algorithm) and Mf (passing the entire excerpts with annotated features for clustering). The results shown per class indicate that there is a high-level of human agreement (ranging from a low of

Table 4.9: Rand index for 5 humans (h1-h5) and machine-generated clusters with (Mf) &amp; without (M) features.

	Permissions						
	h1	h2	h3	h4	h5	M	MF
<b>h1</b>	(1)	0.9	0.65	0.83	0.9	0.86	0.8
<b>h2</b>	***	(1)	0.64	0.93	0.9	0.85	0.89
<b>h3</b>	***	***	(1)	0.65	0.65	0.7	0.72
<b>h4</b>	***	***	***	(1)	0.83	0.91	0.78
<b>h5</b>	***	***	***	***	(1)	0.85	0.88
	Duties						
<b>h1</b>	(1)	0.71	0.89	0.91	0.71	0.93	0.97
<b>h2</b>	***	(1)	0.61	0.63	0.95	0.7	0.68
<b>h3</b>	***	***	(1)	0.92	0.63	0.86	0.86
<b>h4</b>	***	***	***	(1)	0.65	0.85	0.88
<b>h5</b>	***	***	***	***	(1)	0.7	0.67
	Prohibitions						
<b>h1</b>	(1)	0.95	0.75	0.85	0.89	0.87	0.9
<b>h2</b>	***	(1)	0.75	0.85	0.92	0.91	0.92
<b>h3</b>	***	***	(1)	0.73	0.76	0.76	0.75
<b>h4</b>	***	***	***	(1)	0.89	0.8	0.81
<b>h5</b>	***	***	***	***	(1)	0.86	0.86

Table 4.10: Average results.

	Human	M	Mf
<b>Permission</b>	0.79	0.83	0.81
<b>Duty</b>	0.76	0.81	0.81
<b>Prohibition</b>	0.83	0.84	0.85

0.61 to a high of 0.95). There is also a high-level of human-machine agreement (ranging from a low of 0.67 to a high of 0.97). This indicates that in general, there was higher disagreement between humans than between machine and humans. To further help with interpreting the above results, table 4.10 enlists the agreement between the evaluators (Human), between the benchmark method and the evaluators (M) and between the feature-based clustering and the evaluators (Mf); once again organized per class.

The above results indicate that the applied clustering method based on the OBIE-derived instances is relatively prosperous, considering the level of human disagreement when performing the same task manually; but that there is no significant difference between the two clustering methods. On the other hand according to table 4.10, the feature-based results are closer to human agreement. The accumulated deviation of **MF** from **Human** is 9%, whereas the aggregated deviation of **M** from **Human** is 10%. Although the difference is minor, the results reconfirm our previous assumption that the feature-based approach generates more fined-grained clusters and is more attuned to human intuition and perception. While we are encouraged by the former result, we see the potential to further improve the feature-based clustering algorithm. The said features are already being used to improve the results shown in *EULAide* — the cluster titles, shown in figure 4.3, are based on the derived *actions*.



### 4.2.3 Usability Experiments

This section provides the evaluation details of *EULAide* usability from an end-user point of view. This experiment seeks to estimate whether our approach helps users to understand and digest EULAs. First, we will explain the setup for running the experiment, and then we will present the results.

#### Experimental Setup

In order to conduct our evaluation, we chose the same four EULAs from our previous experiment (section 4.2.2) for the current task. A legal expert was asked to design five multiple-choice questions for each EULA (e.g., twenty in total). All questions are related to `permissions`, `prohibitions` and `duties`. Afterward, six people from the university campus were selected to take part in our experiment. Each person studied all four EULAs: they had to read two EULAs in the original text and also exploit the *EULAide* service for the other two EULAs. With this setting, each EULA in the natural text was read by three different students. Similarly, each EULA was browsed and studied with *EULAide* by three individuals. The participants were aware that their tasks are related to `permissions`, `prohibitions` and `duties`, and they were asked to understand and digest EULA in order to answer the questions.

The question-answering phase was split into two stages: first, the participants had to answer the questions using their memory and without looking at the EULA; Second, if they were unsure about some questions, they could check the EULA and use search tools to find the answer. The rationale behind this setting was to measure how good users can remember policies and also how fast they can search for information in the EULA. The primary purpose of *EULAide* is to get an overview of a EULA before accepting it. In practice, when one agrees with terms and conditions, they should try to remember the primary parts of the license agreement in order to avoid infringing the terms.

#### Evaluation & Discussion

In this section, we have reported the average results of our experiment, e.g., for each EULA - either in natural text mode or *EULAide* mode - the average of three values (corresponding to three participants) was computed. In this case, we have eight average values: four EULAs in natural text mode plus the same four EULAs in *EULAide* mode. Finally, for simplicity and avoiding confusion, only the average value of each mode is presented here.

Table 4.11 shows the average time for each step of the experiment. According to the table, using *EULAide* to study and understand EULA takes significantly less time than reading the EULA in natural text, which was indeed an expected result. Furthermore, as already mentioned before, we have divided the answering phase into two steps: **phase1** is based on memory, and **phase2** is allowing the users to search for the answer in the EULA. Not surprisingly, the average times of **phase1** are very similar, because regardless of the mode (natural text or *EULAide*), reading the questions and their multiple choice answers takes relatively equal time. However, regarding **phase2**, using *EULAide* to search for answers is one minute and 15 seconds faster than finding the desired information in the natural text license. Once again, this was an expected outcome, e.g., searching in a structured text is slightly more straightforward.

The second part of the current evaluation is concerned with the correctness of answers provided by the participants. Table 4.12 shows the average percentage of results. According to this table, the correctness of answers in the natural text mode is 5% higher than *EULAide* mode, which is a reasonable result. If the end-user bears with the cumbersome legal lingua in the EULA and spends time studying it, he/she can understand the important parts of the license. However, as stated in the introduction, only a few people read EULA and *EULAide* is an attempt to motivate them to be aware of what they are agreeing to. Our

Table 4.11: Average time (in seconds).

	Reading	Answering Phase1	Answering Phase2
<b>EULA-Full</b>	1185	75	152
<b>EULA-EULAide</b>	315	72	77

Table 4.12: Average percentage of questions results (%).

	Correct	Incorrect	Unanswered in Phase1		
			Phase2 Correct	Phase2 Incorrect	Phase2 Unanswered
<b>EULA-Full</b>	67	8	18.5	5	1.5
<b>EULA-EULAide</b>	62	15	6.5	4.5	12

result shows that if end-users exploit our tool, they can get on average 62% of the questions right, which is indeed very encouraging. Finally, there are some unanswered questions in **phase1**, which leads us to the next phase. In the **phase2** of the answering process, the participants were allowed to search for information in the EULA. According to the table, from 25% unanswered questions in the full-text mode, people have found 18.5% correct answers after re-looking. On the other hand, from 23% unanswered questions in *EULAide* mode, only 6.5% of correct answers were found. Consequently, 12% of unsure questions remained unanswered. This is due to the semi-automatic process of information extraction phase. The F-measure of OBIE pipeline is around 75% and not all of permissions, duties and prohibitions can be extracted with the pipeline. Therefore, not all of the questions were covered in *EULAide*, and the participants could not find the answers.

In summary with *EULAide*, people get 19.5% incorrect answers, while with the full text and search, they get 13.3%, i.e., *EULAide* has on average 6% higher error rate (incorrect). Similarly, there are, on average, 10.5% more unanswered questions with our approach. While we expect that every (semi-)automatic approach leads to a potential information loss, we aim to verify that the error rate of *EULAide*, as well as its unanswered questions, are consequences of automatic extraction and summarization. The information loss of (10.5+6)% by *EULAide* is a reasonable cost for the time gained (Exploiting *EULAide* is on average three times faster than the full text), and the increased incentive for familiarizing oneself with a EULA rather than merely accepting it. Furthermore, it should be stated here that the number of selected EULAs for this type of experiment was a bare minimum. EULAs are time-consuming, and it is challenging to find people who are willing to read and understand them. Due to the number of participants and selected EULAs we acknowledge that this experiment does not seek to make a scientific conclusion but rather, provides an indicative result that is open to interpretation.

### Usability Test

The last task for the participants was filling a usability questionnaire for *EULAide* evaluation. We have reused a widespread form available online<sup>10</sup>. There are thirty questions categorized into four groups: 8 questions for usefulness, 11 for ease of use, 4 for ease of learning, and 7 for satisfaction. There are seven options for each question ranging from 1 (strongly dissatisfied) to 7 (strongly satisfied). Table 4.13 shows the average scores of six participants for each category. The results are surely promising and imply

<sup>10</sup><http://garyperelman.com/quest/quest.cgi?form=USE>



Table 4.13: Average scores of six participants for the usability questionnaire (Max=7).

Usefulness	Ease of Use	Ease of Learning	Satisfaction
6.14	6.11	6.75	6.0

that end-users are quite satisfied using *EULAide*. Furthermore, some participants stated a few points regarding the service. The positive feedbacks include a pleasant and friendly user interface, the fast response time (less than 1 minute), significant time reduction concerning EULA digestion, a summary for each cluster, grouping similar policies, and the ability to expand a specific cluster. The improvements given by participants suggest some interesting ideas for future work. Two participants recommended to include other aspects of EULA in the summary, e.g., what are the agreements between them and the service provider? Three users have said not all of permissions, prohibitions and duties are covered by *EULAide*. Last but not least, almost all participants were pleased with the summarization idea and encouraged us to improve the approach in the future.

### 4.3 Summary

In this Chapter, we presented the *EULAide* system, a holistic approach for the analysis and preparation of end-user license agreement. Our approach includes a comprehensive ontology representing relevant terms, an ontology-based information extraction, a clustering method for the extracted excerpts (permissions, prohibitions and duties) and a Web-based user interface for self-service license analysis. In an era of proliferation of online services, facilitating easy and transparent user knowledge of the terms and conditions as laid out in the license agreements can not be overestimated. The domain of licenses and EULAs is well suited for automated information extraction since the legal language used is more constrained and standardized than arbitrary text content.

In order to evaluate the OBIE pipeline, we created and manually annotated a corpus of 20 well-known licenses. For the gold standard we achieved an Inter-Annotator Agreement (IAA) of 90%, resulting in 193 permissions, 185 prohibitions and 168 duties. The ontology-based background knowledge enables us to achieve relatively high precision and coverage with an overall F-measure of 70-74%, which, in the context of the IAA, proves the feasibility of the approach. Furthermore, our evaluation showed that clustering is effective and significantly reduces the number of relevant terms for users to focus on initially. In addition, the usability experiments proved that *EULAide* can significantly improve human comprehension and that improved feature-based clustering has the potential to further reduce the time required for EULA digestion.

Experimental results provide enough empirical evidence to answer RQ1 & RQ2. *EULAide* is more visual and simpler to digest EULAs and saves around 75% of the time. However, we are aware that it comes at a marginal price of 10.5% loss of valuable information, which is an acceptable trade-off, considering the amount of time saved by users - especially since the full EULAs are not being read by a lot of users. We deem this work to be a significant step forward to make the description of rules and regulations governing online services, software tools, portals, and apps more user friendly.



---

# Analysis of Contractual Agreements using Deep Learning

---

Chapter 4 presented a rule-based approach for interpretation of license agreements. However, in the presence of a reliable dataset, machine learning methods are favored over hand-coded linguistic rules. This Chapter is dedicated to the analysis and presentation of contractual agreements using Machine Learning (ML) and, more specifically, Deep Learning (DL). It addresses research question 1 with the focus on the statistical techniques (vs. rule-based, which was addressed in Chapter 4):

## Research Question 1 (RQ1)

Are text mining techniques able to extract valuable information from contractual agreements?

Our running use-case for this Chapter will be privacy policies. We present *Pripolis*, which performs multi-label classification on a privacy policy's paragraphs. Moreover, having information about paragraphs' classes, we predict five risks (*Expected Use*, *Expected Collection*, *Precise Location*, *Data Retention* and *Children Privacy*) and different colour-coded risk levels. The contributions of this Chapter in general are:

- Presentation of a strong baseline for privacy policy classification using Neural Networks (NN) that successfully reproduces state-of-the-art findings and further improves the results by employing the BERT framework [13] for two different gold standards;
- Prediction of risk colors (green, yellow or red) for the five risk icons based on the extracted information;
- A comprehensive set of experiments.

Chapter 5 is based on the following publications [102, 103]. The Chapter is structured as follows: in Section 5.1 we introduce the OPP-115 dataset that will be our core resource for training and testing; Section 5.2 presents a strong baseline for privacy policy classification; Section 5.3 explores an approach for predicting privacy icons' colors, and finally, Section 5.5 examines the research question and provides concluding remarks with an outlook to future work.

## 5.1 Background: OPP-115 Dataset

Within the field of machine learning, there are two main types of tasks: supervised and unsupervised. Unsupervised learning does not require an annotated dataset and infers the natural structure hidden in the data resource without having explicitly-provided labels. On the other hand, the goal of supervised learning is to approximate the relationship between input and output (labels). In the presence of a reliable labeled dataset, supervised learning is used.

A prominent group on the classification and analysis of privacy policies is the *Usable Privacy Policy Project*<sup>1</sup>, who provided OPP-115, the first comprehensive dataset with fine-grained annotations on paragraph level [59]. The project aims to extract valuable information for the benefit of regular and expert end-users. In order to create a dataset, first, they compiled a corpus containing 115 privacy policies from 115 US companies. Then, a small group of domain experts identified different data practice categories and their descriptive attributes from multiple privacy policies through an iterative refinement process. After finalizing categories and attributes, each privacy policy was randomly assigned to 3 experts (out of 10), and fine-grained annotations were created for the whole corpus. Final annotations are in two levels: 10 high-level categories and 24 low-level attributes. The high-level categories are:

1. *First Party Collection/Use*: how and why the information is collected.
2. *Third Party Sharing/Collection*: how the information may be used or collected by third parties.
3. *User Choice/Control*: choices and controls available to users.
4. *User Access/Edit/Deletion*: if users can modify their information and how.
5. *Data Retention*: how long the information is stored.
6. *Data Security*: how are users' data secured.
7. *Policy Change*: if the service provider will change their policy and how the users are informed.
8. *Do Not Track*: if and how Do Not Track signal is honored.
9. *International/Specific Audiences*: practices that target a specific group of users (e.g., Children, Europeans, Californians, Citizens from other countries)
10. *Other*: additional practices not covered by the other categories.

Figure 5.1 demonstrates the hierarchy of the dataset [67]. The top level defines ten high-level classes, and the lower levels demonstrate low-level attributes. For instance, the high-level category *First Party Collection & Use* has 9 low-level attributes. We provided a few examples of attribute values. As shown in the picture, some low-level attributes belong to multiple high-level categories (*Personal Information Type, Purpose, ...*). An individual data practice belongs to one of the ten categories above, and it is articulated by a category-specific set of attributes. Figure 5.2 shows a screenshot of the web-based tool for the expert's annotations, developed by the dataset creators<sup>2</sup>. The selected paragraph in the picture is labeled with two high-level classes: *First Party Collection/Use* and *Third Party Sharing/Collection*. In addition to specifying the high-level categories for each paragraph, the annotators identified attribute values belonging to that specific category. When applicable, they also specified the text spans related to the

---

<sup>1</sup><https://usableprivacy.org/>

<sup>2</sup><https://explore.usableprivacy.org/?view=human>

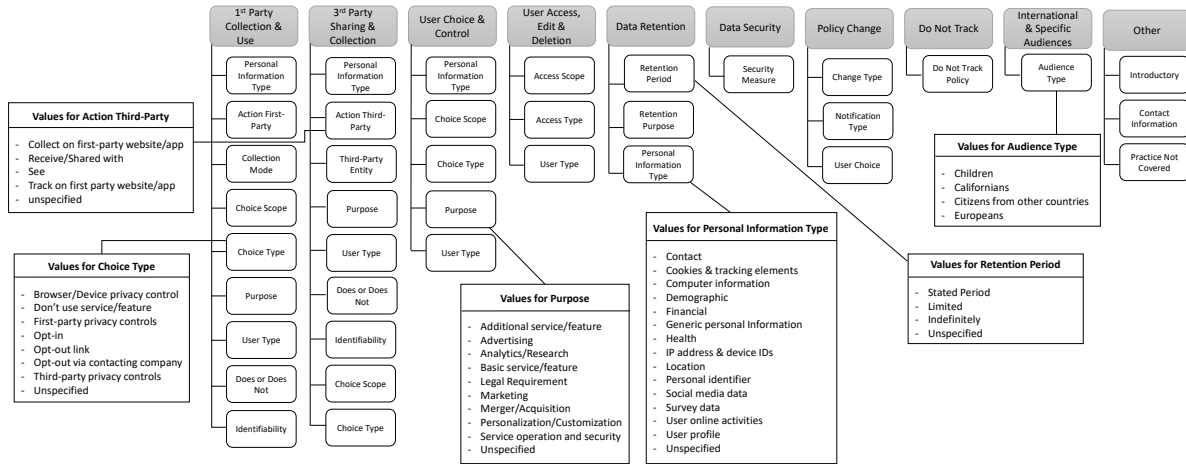


Figure 5.1: The OPP-115 dataset. The top level of the hierarchy (shaded blocks) defines high-level categories. The lower level defines a set of privacy attributes, each assuming a set of values.

The screenshot shows the usableprivacy website interface. At the top, there is a navigation bar with 'User Profile', 'Task', 'Visualize', 'Settings', and 'Logout'. Below this, the current policy is identified as 'a\_98\_neworleansonline.com'. The interface features several tabs for different policy sections: 'First Party Collection/Use', 'Third Party Sharing/Collection', 'User Choice/Control', 'User Access, Edit and Deletion', 'Data Retention', 'Data Security', 'Policy Change', 'Do Not Track', 'International and Specific Audiences', and 'Other'. The 'First Party Collection/Use' tab is active, showing a list of attributes with dropdown menus for their values: Does/Does Not, Collection Mode, Action First-Party, Identifiability, Personal Information Type, Purpose, User Type, Choice Type, and Choice Scope. A paragraph of text is displayed with several words highlighted in yellow and blue, indicating annotations. Below the text, there is a 'Practices of this paragraph' section listing the identified practices: 'First Party Collection/Use' and 'Third Party Sharing/Collection'. The 'First Party Collection/Use' practice is further detailed with a list of values: 'Does Unspecified Collect on website Identifiable Contact Unspecified Unspecified Unspecified Unspecified'. A 'Save' button is visible at the bottom right.

Figure 5.2: An example of annotations by an expert.

attribute values. For instance, according to one of the annotators, the low-level annotations for the *First Party Collection/Use* in the figure are: the website **Does** collect information; the *Action First-Party* value is **Collect on website**; the *Identifiability* of information is **Identifiable**; the collected *Personal Information Type* is the user's **Contact**; and the rest of attributes are labeled as **Unspecified**. This Final dataset consists of 23K data practices, 128K practice attributes, and 103K annotated text spans

OPP-115 comprises 3 792 paragraphs where each paragraph was labeled with one or more high-level classes and low-level attributes. According to the dataset creators, the best agreement was achieved on *Do Not Track* class with Fleiss' Kappa equal to 91%, whereas the most controversial class was *Other*, with only 49% of agreement [59]. The latter category was further decomposed into its attributes: *Introductory/Generic*, *Privacy Contact Information* and *Practice Not Covered*. Therefore, in the case of high-level categories, we face a multi-label classification problem with 12 classes. It should be clarified here that computing Fleiss' kappa considering all categories together is not feasible for OPP-115, as annotators differ per policy. Aforementioned, there were ten experts, and each policy was randomly assigned to 3 of them. If three experts were the same experts for the whole dataset, it was rational to compute an overall Fleiss's kappa for all ten categories and between 3 annotators. For this reason, [59] reported Fleiss' kappa per category.

Along with the original dataset, the group released three consolidated sets regarding the low-level attributes. Their consolidation procedure merges the experts' annotations if the annotations refer to the same underlying data practices in the text. In order to perform the consolidation, a list of requirements was identified:

- the selection of data practices are eligible to be merged if they belong to the same category;
- at least two annotators contributed to the selected data practices;
- the selected data practices belong to the same segment (paragraph).

After finalizing the requirement list, the different eligible combination of data practices was created, scored, ranked, pruned with a threshold, and finally consolidated based on the ranks until no further consolidation was possible. In summary, the procedure is as follows:

1. the consolidation sets which include three experts' annotations have priority over sets containing two;
2. the data practices in a specific consolidation set is replaced by a "master" data practice where the attribute values are merged based on the majority-vote if possible, otherwise is set to *Unspecified*;
3. when creating the "master" data practices, the text span associated with each practice is produced with a strong bias toward recall, i.e., creating a new text span that begins and ends with the first and last indexes in the set.

The scoring method is based on the summative overlap between the sets of text spans associated with attributes, meaning that the score for two data practices with a high text span overlap is high, and the score for two practices that are associated with different text is low. Finally, several threshold values are applied to create consolidated sets. The three released datasets have the threshold values of 0.5 0.75, and 1. Generally, the average number of practices produced by consolidation is less than the average practices per annotator per segment (2.04). We chose the dataset with the threshold value of 0.5, since it contains the most annotations. In the following sections, we explain our approach founded on the OPP-115 dataset.

## 5.2 Establishing a Baseline for Privacy Policy Classification

Various studies indicate that, despite their proliferation, a majority of consumers still skip privacy policy consent forms due to the difficulty required for lay users to comprehend their contents [2, 5, 104]. To assist end-users with consciously agreeing to the conditions, we consider deep learning to classify privacy policy paragraphs into pre-defined categories for easier comprehension. Our efforts seek to build on the results of two earlier dominant studies in the literature. The first is the OPP-115 dataset, which was introduced in the previous Section. The second study which inspired our research is the effort by *Polisis* to build a Convolutional Neural Network (CNN) model exploiting OPP-115 [67]. As addressed in Section 3.2, despite the valuable contribution of these earlier studies, they exhibit one major weakness: reproducibility. Due to a lack of information on the exact ML dataset splits used, and the lack of a universal gold standard in the literature, subsequent studies have created their own. This makes it difficult to interpret and compare the different results collectively. A significant contribution of the efforts presented here is our provision of a robust and utterly reproducible baseline for future research.

For high-level categories, we first compiled two gold standards from OPP-115: one based on majority votes (i.e., two or more experts agree on a label), and the other with the union of all expert annotations. The dataset creators [59] considered the majority-vote-based standard, whereas *Polisis* used the union-based, with the rationale that disagreements are a result of the experts' high understanding of legal texts and that therefore, none of their annotations should be deemed incorrect. In order to establish a firm foundation, we compare three models with both gold standards. The first model is a CNN using word embeddings, whose generation is directly comparable to the earlier *Polisis* efforts. The second and third models are based on Bidirectional Encoder Representations from Transformers (*BERT*) [13] that has state-of-the-art performance on many NLP tasks. The results attained demonstrate consistency and a significant improvement over the baseline and indicate good reliability: A 77% micro-average F1 on the union-based gold standard, and a 85% micro-average F1 on the majority-based gold standard.

For the low-level attributes, we perform our experiments with the fine-tuned *BERT* model, since it shows the best performance for the high-level categories. The dataset, in this case, is a consolidated set provided by the dataset creators. In the consolidation process, the annotations from multiple experts are merged together if those annotations refer to the same data practices in the text. In Section 5.4.1, a detailed explanation of experiments is presented.

### 5.2.1 Pre-trained Word Embeddings

Traditionally, text classifiers have taken advantage of vector representations like bag of words or term-frequency inverse-document-frequency (TF-IDF). However, this method has the disadvantage of not retaining the semantic information depicted by the order of words, as well as the meaning of the single words as independent units and be purely dependent on the context. Thus, we investigate word embeddings.

Word embeddings were initially proposed by [105, 106] and were later popularized by [107]. The continuous bag of words (cbow) method, which is a variant of word2vec, creates a numeric representation of words by attempting to predict a given word by considering its neighbors as seen in text. A huge benefit such an algorithm is that no labeled data is necessary, but only great amounts of correct text.

While word2vec is effective at storing some semantic meaning in a vector representation, it treats words as atomic units, and thus, it does not take into consideration the internal structure of words. Such information can be useful for less frequent words or compound words like rainfall or greenhouse. FastText uses a bag of character n-grams to represent words, where each character n-gram is a vector, and all the constituents are summed up to create a representation for the word [108, 109].

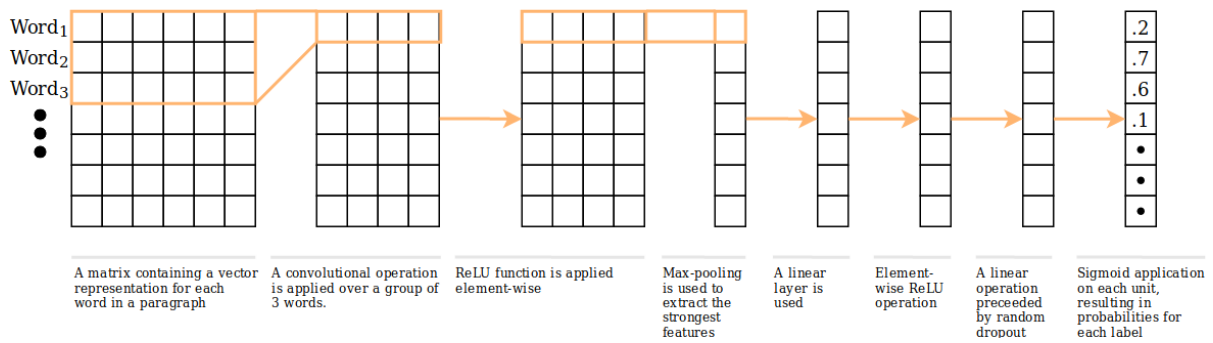


Figure 5.3: CNN architecture for multi-label classification of privacy policies.

The aforementioned properties can be useful for the context of privacy policies. Since most openly available word embeddings are trained on the news or Wikipedia corpora [109], we utilize fastText to create vector representations that are more suitable for the current task. For that purpose, we used a big corpus of 130k privacy policies scraped from an application store for smartphones. In app stores, applications are required to provide privacy policies. After tokenizing the text with NLTK [110], there are 132 595 084 tokens in total, and 173 588 unique ones. We compared the vocabulary between this corpus and two versions of OPP-115 that we utilize (majority-vote & union). We saw that there are 1 072 words which are seen only in OPP-115 majority-vote version, but not in the corpus used for drafting the word vectors. Similarly, for the gold standard containing the union of all classes, there were 1 119 out-of-vocabulary (OOV) words. The difference in the amount of OOVs is due to the fact that the majority vote dataset has fewer paragraphs (when there was no agreement on a single category) than the union-based, and thus, it is less likely that there are unseen words. More details regarding the size of the dataset versions are provided in Section 5.4.1. After manual inspection, we concluded that most of the out-of-vocabulary words are names of brands, products, services, or their web addresses. These are entirely omitted since, from an intuitive perspective, they should not be decisive for the correct detection of a policy class. Hence, the vocabulary is sufficient for the task.

## 5.2.2 Convolutional Neural Network

To tackle the multi-label classification problem, we follow the work of [67] by using a CNN (displayed in Figure 5.3). The previously explained word embeddings are provided as input to the neural network. A convolutional operation is applied with a context window of 3 words, whose output then passes through a Rectified Linear Activation (ReLU) function. Then, from each context output, only the strongest features are selected by a max-pooling layer, resulting in a single vector that contains the most informative properties of each context. Thus the neural network is forced to focus only on certain features that are specific to the current goal. Furthermore, a linear layer followed by a ReLU is applied to create a higher level representation of the collected information. Finally, a linear layer with as many nodes as classes is applied to provide an output in the target dimensions and passed through a sigmoid function to obtain per label probability scores.

The proposed architecture shares a strong resemblance with the work of [10], where a CNN is used for multi-class classification of sentences. However, it lacks a random dropout just before the last linear layer. We conduct experiments with 50% dropout. Additionally, we used Adam [111] optimization algorithm combined with early stopping. The convolutional neural network is optimized using binary cross entropy



loss:

$$\ell(x, y) = L = \{l_1, \dots, l_N\}^T \quad (5.1)$$

$$l_n = -w_n [y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)] \quad (5.2)$$

where  $l_1, \dots, l_N$  specify the 12 loss values for each of the 12 possible labels that we have in the dataset. It is being calculated for each since this is a multi-label classification, and we could have any combinations of those. After we have the 12 losses, we take the mean of those 12 to get one scalar number. Furthermore,  $x$  is the model prediction,  $y$  is the true label,  $w$  is the class-specific weight which in our case are all 1. For instance, if we consider that our current model assigns probability  $p$  to observation  $o$  for the *Data Retention* label, the loss function for this specific label will be:

$$\text{loss}(\text{DataRetention}) = y \cdot \log p + (1 - y) \cdot \log(1 - p) \quad (5.3)$$

where  $y$  is 1 if observation  $o$  is labeled with *Data Retention* in the gold standard and 0 if not.

### 5.2.3 Bidirectional Encoder Representations from Transformers

The *BERT* framework [13] uses several layers of transformer encoders [112] to create a bidirectional representation of the tokens in the sequence. The approach operates in two stages: first, the model is pre-trained on large amounts of unlabelled data; second, it is fine-tuned on specific labeled data to solve a downstream problem, which in our case is multi-label classification.

To handle various domains and tasks, *BERT* is using WordPiece [113] tokenization. It provides a reasonable balance between character and subword level information. For example, a model using it, can detect similar suffixes or roots among words. This way, the vocabulary stays within a reasonable size, without having too many entries. The chosen vocabulary size is 30 000 [13].

*BERT* is pre-trained using two unsupervised tasks. The first one is masked language modeling (MLM), i.e., the model is being taught to predict 15% of the randomly “masked” tokens in a sentence. The masking uses one of three randomly chosen possible ways: 1) in 80% of the cases, a token is replaced with [MASK]; 2) in 10% with another random word; and 3) in the remaining 10% no replacement is done [13]. The other unsupervised language modeling task is next sentence prediction (NSP). Every input sequence to the framework always starts with the classification token [CLS], which provides a fixed-length representation for the whole input. For NSP, two subsequent sentences from the corpora are concatenated with another separator token, [SEP], so that the model is aware of the separation between the two. In 50% of the cases, the second sentence is replaced by another one. Thus, *BERT* is trained to recognize when a pair of sentences appear together in the corpora (or they do not), using the classification token [13].

We use a pre-trained version of *BERT*<sub>BASE</sub><sup>3, 4</sup> which has 12 encoder layers, a hidden state size of 768, and 12 attention heads, totaling in 110M parameters. Additionally, we also prepare another fine-tuned version of the language model with our 130K privacy policy corpus<sup>5</sup>. Ninety percent of those were used for training while the remaining ten for validation. We fine-tuned the model for three epochs and achieved a cross-entropy loss on the mask language model task of 0.1151 and perplexity, 1.1220. Finally, both versions of the approach are trained for the high-level classification task and then the best performing

<sup>3</sup><https://github.com/huggingface/transformers>

<sup>4</sup><https://github.com/kaushaltrivedi/fast-bert>

<sup>5</sup>The BertLMDataBunch class contains from\_raw\_corpus method that takes a list of raw texts and creates DataBunch for the language model learner.

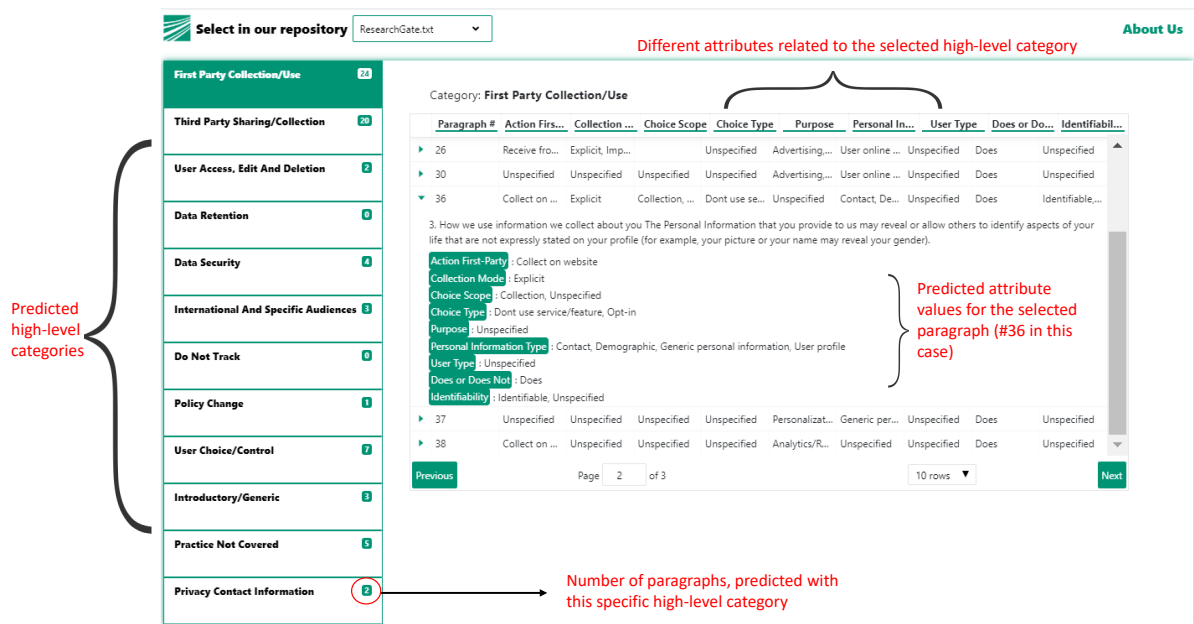


Figure 5.4: Pripolis Web interface showing the predicted (high/low)-level classes for a user provided privacy policy.

model is applied for the low-level attribute classification. For more details on *BERT*, we refer the reader to the relevant references [13, 112].

### 5.2.4 Pripolis Framework

We developed a Web interface for a structured presentation of classes predicted by *Pripolis*. Figure 5.4 demonstrates an example of the results for the *ResearchGate* privacy policy. The user can either choose one of the privacy policies in our repository or copy/paste the text from a target website. Using line breaks, a simple script divides the policy’s text into paragraphs. Then for all paragraphs, high-level categories will be predicted. Afterward, based on the predicted category and the hierarchy in 5.1, the corresponding attribute models are called, and the attribute values are predicted. Finally, all the predicted information will be presented to the user. The left pane shows all the predicted high-level categories along with the number of paragraphs related to that class. The user can choose the category they are interested in and examine all segments that are predicted with the desired category. Furthermore, it is possible to expand each segment and view detailed information regarding that segment.






## 5.3 Risk Level Prediction

In light of the, now enforced EU-wide, General Data Protection Regulation , there has been an increasing interest in privacy policy analysis as this new set of regulations increases the constrains for companies holding customers data [19–21, 73]. As addressed in 3.2, some studies attempted to measure risk levels for a set of pre-defined factors [62, 63, 67]. In this Section, we use the fine-tuned *BERT*’s predictions and predict risk levels (i.e., Green, Yellow, or Red) for five privacy icons.

*Disconnect*<sup>6</sup> is a tool that aims to inform end-users about their personal data usage when browsing online services and Websites. The tool is available as a browser extension and mobile app. One of the

<sup>6</sup><https://disconnect.me/>

Table 5.1: *Disconnect* privacy icons with their descriptions & *Polisis*'s interpretation from Harkous et. al. [67].

Icon	Disconnect Description	Disconnect Color	Assignment	<i>Polisis</i> ' Interpretation as Labels	<i>Polisis</i> 's Automatic Color Assignment
	Discloses whether it allows other companies like ad providers and analytics firms to track users on the site?	<b>Red</b> <b>Yellow</b> <b>Green</b>	Yes, w/o choice to opt-out. Or, undisclosed. Yes, with choice to opt-out. No.	Let $S$ be the segments with category: <i>Third Party Sharing &amp; Collection</i> , <b>purpose</b> : $\in$ [advertising, analytics-research ], and <b>action-third-party</b> $\in$ [track-on-first-party-website-app, collection-first-party-website-app].	<b>Yellow</b> All segments in $S$ have category: <i>User Choice/Control</i> and <b>choice-type</b> $\in$ [opt-in, opt-out-link, opt-out-via-contacting-company] <b>Green</b> $S = \emptyset$ <b>Red</b> Otherwise
	Discloses whether the site or service tracks a user's actual geolocation?	<b>Red</b> <b>Yellow</b> <b>Green</b>	Yes, possibly w/o choice. Yes, with choice. No.	Let $S$ be the segments with <b>personal-information-type</b> : location.	
	Discloses whether data it collects about you is used in ways other than you would reasonably expect given the site's service?	<b>Red</b> <b>Yellow</b> <b>Green</b>	Yes, w/o choice to opt-out. Or, undisclosed. Yes, with choice to opt-out. No.	Let $S$ be the segments with category: <i>First Party Collection &amp; Use</i> and <b>purpose</b> : advertising.	
	Discloses how long they retain your personal data?	<b>Red</b> <b>Yellow</b> <b>Green</b>	No data retention policy. 12+ months. 0-12 months.	Let $S$ be the segments with category: data-retention.	<b>Green</b> All segments in $S$ have <b>retention-period</b> : $\in$ [stated-period, limited] <b>Red</b> $S = \emptyset$ <b>Yellow</b> Otherwise
	Has this website received TrustArc's Children's Privacy Certification?	<b>Green</b> Gray	Yes. No.	Let $S$ be the segments with category: <i>International &amp; Specific Audiences</i> and <b>audience-type</b> : children	<b>Green</b> length ( $S > 0$ ) <b>Red</b> Otherwise

project's highlights is the definition of privacy factors. *Polisis* has founded the risk color assignment on *Disconnect* icons. Table 5.1 shows five *Disconnect*'s privacy icons, their descriptions and *Polisis*'s interpretation of icons. In this table, the high-level categories are in *italic* and the low-level attributes are in **bold**. As a proof-of-concept, we will use table 5.1 for our risk color prediction.

## 5.4 Experimental Study

*Polisis* exploits deep learning to extract pre-defined labels, classes, and values from privacy policies and predicts risk colors for a set of privacy icons. In this Section, we answer the first research question by evaluating the classification performance as well as measuring the risk color prediction accuracy.

### 5.4.1 Multi-label Classification Evaluation

In pursuance of providing a reliable baseline for privacy policy classification, two gold standards were compiled for the high-level categories: union-based, which contains all expert annotations; and the majority-vote-based gold standard, where only annotations with an agreement between at least two experts were retained. Label distributions in both gold standards are shown in table 5.2. Following conventional ML practices, dataset splits are randomly partitioned into a ratio of 3:1:1 for training, validation, and testing, respectively; while maintaining a stratified set of labels. In total, the union-based dataset contains 3 788 unique segments, and the majority-based one comprises 3 571 unique segments<sup>7</sup>. The latter has fewer segments due to the 217 paragraphs that were eliminated because no expert agreement was reached.

In the case of multi-label classification, it is not clear which average (macro or micro) best defines a model's performance. As Sebastiani argues, there is no agreement to choose between micro- and

<sup>7</sup>All splits are available for further experiments.

Labels	Majority-vote gold standard					
	Tr	V	T	Tr(%)	V(%)	T(%)
First Party Collection & Use	781	176	250	34.2	30.9	35
Third Party Sharing & Collection	584	158	203	25.5	27.7	28.4
User Access, Edit and Deletion	101	24	24	4.4	4.2	3.4
Data Retention	50	14	14	2.2	2.4	2
Data Security	139	31	40	6.1	5.4	5.6
International/Specific Audiences	204	41	56	9	7.2	7.8
Do Not Track	22	6	3	1	1	0.4
Policy Change	73	25	21	3.2	4.4	3
User Choice/Control	233	48	77	10.2	8.4	10.8
Introductory/Generic	240	72	78	10.5	12.6	11
Practice Not Covered	83	21	25	3.6	3.7	3.5
Privacy Contact Information	129	32	42	5.6	5.6	5.9
	Union-based gold standard					
First Party Collection & Use	988	243	288	40.8	40.1	38
Third Party Sharing & Collection	755	204	227	31.1	33.7	30
User Access, Edit and Deletion	155	29	46	6.4	4.8	6.1
Data Retention	111	21	24	4.6	3.5	3.2
Data Security	251	65	59	10.3	10.7	7.8
International/Specific Audiences	225	67	61	9.3	11.1	8.1
Do Not Track	22	3	7	1	0.5	0.9
Policy Change	118	27	47	4.9	4.4	6.2
User Choice/Control	405	97	130	16.7	16	17.2
Introductory/Generic	514	137	162	21.2	22.6	21.4
Practice Not Covered	402	102	138	16.6	16.8	18.2
Privacy Contact Information	207	44	72	8.5	7.3	9.5

Table 5.2: Label distribution in the two gold standards for the high-level categories; Tr:Train; V:Validation; T:Test.

macro-averages in literature [114]. Some studies claim that macro-average is fair in case of class imbalance, since all the categories have the same weight, whereas micro-average favors methods that just correctly predict the most frequent categories [115]. However, others (the majority) believe that when the label distribution is not balanced, computation of micro-average is preferable, because micro-average aggregates the contributions of all classes to compute the average metric [101, 116]. In order to establish a firm foundation, we report both averages.

Table 5.3 presents F1 scores across high-level categories with a threshold equal to 0.5 for the two gold standards. For CNN, we applied Adam with default parameters and with 50% dropout just before the last linear layer (learning rate = 0.001, decay rates:  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ). *BERT* is optimized with the default configuration and LAMB optimizer [117].

In total, six experiments were carried out for the high-level classification. The scores obtained (micro-averages ranging from 70-85% and macro-average in range of 65-76% for both gold standards) are considered very accurate, especially in the context of the Fleiss expert agreements, reported in [59], which showed human agreement between 49-91% for the same classes here considered. As expected, for all six experiments, micro- outperform macro-averages, because, for a few labels, the model is not able to learn the class weights properly due to sample scarcity. For instance, *Data Retention* corresponds

Labels	Majority-vote gold standard					
	CNN		BERT		BERT-fine-tuned	
	V	T	V	T	V	T
First Party Collection/Use	83	82	87	88	88	91
Third Party Sharing/Collection	84	82	86	85	87	90
User Access, Edit & Deletion	80	70	82	63	77	73
Data Retention	43	40	42	33	54	56
Data Security	76	75	87	82	87	80
International/Specific Audiences	96	82	94	81	95	83
Do Not Track	91	100	80	100	80	100
Policy Change	80	88	80	88	85	90
User Choice & Control	77	72	75	81	78	81
Introductory/Generic	63	73	75	76	78	79
Practice Not Covered	8	13	18	32	35	35
Privacy Contact Information	86	84	79	80	79	78
Macro Averages	72	71	74	74	77	<b>79</b>
Micro Averages	79	78	81	82	83	<b>85</b>
Union-based gold standard						
First Party Collection/Use	83	81	83	84	87	86
Third Party Sharing/Collection	80	79	79	82	83	86
User Access, Edit & Deletion	56	45	54	49	56	65
Data Retention	36	48	36	68	62	71
Data Security	66	72	71	80	73	76
International/Specific Audiences	89	92	87	93	92	92
Do Not Track	80	60	80	60	100	92
Policy Change	69	77	75	78	77	80
User Choice & Control	66	64	64	63	66	65
Introductory/Generic	63	65	74	68	73	67
Practice Not Covered	41	37	44	46	45	48
Privacy Contact Information	79	71	75	71	83	78
Macro Averages	67	65	68	70	75	<b>76</b>
Micro Averages	72	70	73	74	77	<b>77</b>

Table 5.3: F1 for three models on the two gold standards in (%) with tuned epochs on validation; Threshold=0.5; V:Validation; T:Test.

to only 2-3% of the dataset, and yet this class has 1/12 weight in macro-average calculation; whereas micro-average considers dataset heterogeneity and decreases the impact of scarce categories on the final result. Furthermore, the category *Practice Not Covered* shows low F1 on both gold standards. This category refers to all practices that are not covered by the other 11 categories and therefore represent a broad range of topics. Consequently, due to the diversity of vocabulary, it is difficult for the model to learn this specific class.

Table 5.3 shows that even  $BERT_{BASE}$  achieves state-of-the-art and further improves the results (without domain-specific embeddings). This is due to the facts that 1) transformers scale much better on longer text sequences because they operate in a concurrent manner; 2)  $BERT$  is using WordPiece encoding and therefore it has a dictionary which is hard to have an OOV case with it; and 3) it has been trained

Measure	CNN		BERT		BERT-fine-tuned	
	V	T	V	T	V	T
Precision	81	81	81	84	81	83
Precision-absence	94	94	94	95	95	95
average	86	86	86	89	88	89
Recall	58	57	60	62	70	71
Recall-absence	97	97	97	97	97	97
average	78	77	79	80	84	84
F1	67	65	68	70	75	76
F1-absence	95	95	95	96	96	96
average	81	80	82	83	86	86

Table 5.4: Macro averages on the union-based gold standard in (%) with tuned epochs on validation; Threshold=0.5; V:Validation; T:Test.

on massive amounts of data. Moreover, the fine-tuned  $BERT_{BASE}$  with 130K corpus privacy policy has significantly enhanced F1 average on both gold standards<sup>8</sup>. Interestingly, fine-tuned  $BERT$  has improved macro-average more than micro. It proves that exploiting a good language model enables the classification model to learn the weights more properly, even with the scarce number of samples.

In order to compare our result to *Polisis*, we present table 5.4 which provides macro-averages on the union-based gold standard. As mentioned in section 3.2, *Polisis* used the union-based dataset to report their results. The average lines in the table represent the macro-average of the metric (precision, recall or F1) in predicting the presence of each label and predicting its absence (the 7<sup>th</sup> line in the table - F1 - is also included in table 5.3).

As shown in table 5.4, we successfully reproduce *Polisis* findings (although with different splits, which remain unavailable) and further improve the result by 5% compared to the state-of-the-art. However, we believe this type of average is not a fair measure for multi-label classification. Table 5.3 shows that the fine-tuned  $BERT$  model has nevertheless significantly enhanced macro-averages (from 65% to 76%) which is not visible in table 5.4, where the enhancement is limited to 5%.

Regarding the low-level classification, we conducted our experiments with the model that yields the best performance for the high-level categories, e.g., the fine-tuned BERT. The low-level classification process shares a strong resemblance with the high-level classification. The only difference lies in the dataset usage, where we utilized the consolidated set with the threshold value of 0.5. In total, 21 classifiers were trained for the low-level attributes. The full results of attribute classifiers are presented in Appendix A.

## Discussion

Our proposed baseline considers notoriously cumbersome privacy policies and investigates automatic methods to assist end-users in comprehending these contractual agreements. The conducted experiments confirm the feasibility of our approach in reaching this objective. Since we are benefiting from supervised ML, the performance of the generated model highly depends on the training dataset quality. As shown in table 5.2, there is a huge difference between the two gold standards for the *Practice Not Covered* class. In the union-based dataset, 642 segments are categorized as *Practice Not Covered*, whereas the majority-

<sup>8</sup>Fine-tuning BERT took 33 hours for 3 epochs on a single GPU. Once it is completed, training the classification model takes only a few hours, depending on the number of epochs.



“ [...] Amazon.com does not sell products for purchase by children. We sell children’s products for purchase by adults. If you are under 18, you may use Amazon.com only with the involvement of a parent or guardian. [...] ”

- *International and Specific Audiences*
- *Practice Not Covered*

Figure 5.5: Disagreement example for the Amazon privacy notice.

based gold standard only records 129 occurrences. Unsurprisingly, for this specific label, all models trained with the union-based dataset outperform the models which were trained by the majority-based one. In addition, 513 variation for the *Practice Not Covered* category between the two gold standards shows high expert disagreement. This was not evident in the original paper [59], because the authors reported Fleiss’ Kappa on the parent category (*Other*) and there is no information on annotator agreement for its subcategories.

Figure 5.5 shows an example of disagreement on *Practice Not Covered* category in the two gold standards. The shown paragraph explains Amazon’s policy on treating children’s data. In the union-based dataset, this segment is annotated with *International and Specific Audiences* and *Practice Not Covered* classes, whereas, in the majority-based, it is only labeled with *International and Specific Audiences*.

Regarding label-specific performance, almost all models perform quite well on *Do Not Track* class in spite of the low sample occurrence. This is probably due to a smaller set of terminology that is often used in such paragraphs, including specifically the word *track*. Furthermore, as mentioned earlier, the best human agreement was also achieved on *Do Not Track* class with Fleiss’ Kappa equal to 91%, which indicates that our ML models simulate human thinking fairly.

The *BERT* model proves that a good language model achieves high performance even on a domain-specific dataset. It also shows that there is a huge potential to improve the results by fine-tuning the language model with domain vocabularies.

In summary, OPP-115 has proven to be a small, yet reliable dataset for supervised privacy policy classification. However, our experiments confirmed legal text subjectivity for a few classes. One possible solution is decomposing those categories into less controversial subclasses with higher experts agreement. In the above example (Figure 5.5), breaking the *Specific Audiences* segment into more specific classes will make annotations less subjective, for human experts and machines alike.

## 5.4.2 Risk Icons Evaluation

Given than we achieve promising F-measure in the multi-label classification of high-level categories and low-level attribute values, it is intuitive to consider predicting risk levels for a set of pre-defined factors. In order to conduct the evaluation, first, we merged the validation and test splits from our majority-vote gold standards. Then, we produce risk colors according to the experts’ annotations and *Polisis* interpretation (Table 5.1). Therefore, our final risk gold standard has five new columns corresponding to the five privacy icons. It is worth mentioning that some segments with multiple labels correspond to different icons’ interpretation. In this case, we retain the color that has a higher risk (Red > Yellow > Green).

After creating the gold standard for risk evaluation, we ran the fine-tuned *BERT* model trained with the majority-vote dataset. Running the model produces high-level categories along with low-level attribute values. Next, we executed the risk color rules presented in the rightmost column of Table 5.1 and computed the accuracy compared to our risk gold standard. Table 5.5 presents the results. As shown in the table, *Polisis*’s test set contains 50 privacy policies (out of 115) where our test split is on the paragraph

Table 5.5: Accuracy of risk color prediction for the five privacy icons; R:Red; G:Green; Y:Yellow.

Icons	Polisis				Pripolis			
	Acc.	Nr (R)	Nr (G)	Nr (Y)	Acc.	Nr (R)	Nr (G)	Nr (Y)
Precise Location	0.84	32	14	4	0.98	61	948	3
Expected Use	0.92	48	8	1	0.96	52	961	0
Expected Collection	0.88	35	12	3	0.97	41	969	2
Data Retention	0.8	29	16	5	0.98	980	18	15
Children Privacy	0.98	12	38	NA	0.99	942	72	NA

level, rather than the whole policy. Consequently, the reported accuracy by *Polisis* is the fraction of policies where the icon based on automatic labels matched the icon based on the experts' labels, where our accuracy is based on the fraction of segments. However, in spite of different data splits, our results outperform state-of-the-art and encourage us to pursue risk level prediction with the assistance of legal experts.

In the light of recently enforced data protection laws in the EU, all parties that use and collect personal information must ensure their compliance with GDPR. Although OPP-115 consists of policies defined by American companies, most of the top-level categories can still be largely mapped to GDPR articles. For instance, the category *First Party Collection/Use* can reflect many practices stated in the Article 13, 'Information to be provided where personal data are collected' and *User Access, Edit & Deletion* can be linked to Articles 16 & 17 ('Right to Rectification/Erasure')<sup>9</sup>. The approach presented here is a valuable initial step towards compliance checking of privacy policies.

## 5.5 Summary

In this Chapter, we investigated the potential of automatic classification of contractual agreements in privacy policy consent forms that are frequently faced by lay users. Our findings are based on the compilation of two gold standards, thus providing a reference privacy policy classification baseline for the relevant research community. To the best of our knowledge, this is the first effort towards a standardized benchmark for privacy policies experiments.

Experimental results provide enough empirical evidence to answer RQ1 with a focus on statistical techniques. In the case of high-level classification, the fine-tuned *BERT* yields F1 score highs of 77-85% (micro-avg) and 76-79% (macro-avg) for union-based and majority-vote-based gold standards, respectively. Both metrics outperform the reported state-of-the-art. In light of human annotator agreement levels achieved for the same data and classes (ranging from 49%-91%), the results can safely be considered as successful. Furthermore, our risk level prediction evaluation with the average accuracy of 97.6%, supports the potential of automatically assigning risk colors to privacy icons.

The approach and method presented are completely reproducible and all resources and data splits are openly accessible<sup>10</sup>. Since the context surrounding our methods (including the data splits) are available, they can be used as a benchmark for other approaches exploring machine-assisted privacy policy classification for improved human understanding.

In conclusion, we intend to continue building upon the baseline achieved and the positive results

<sup>9</sup>Website privacy policies in European union depend also on Directive 2002/58/CE.

<sup>10</sup>A *supplementary archive* is available online for download: <[https://github.com/SmartDataAnalytics/Polisis\\_Benchmark](https://github.com/SmartDataAnalytics/Polisis_Benchmark)>. The archive contains *inter alia* the source-code required to reproduce all the experiments, some useful documentation and necessary datasets.



presented in this Chapter. As demonstrated by the EU-wide GDPR implementation, data regulation is increasingly recognized as a critical area at a political and governance level, whose impact is felt by all digitally-enabled world citizens. Therefore, although not novel, the application of AI techniques to this area has renewed relevance, and there is great value in exploring automation to support private users entering contractual agreements to have a clearer and more secure understanding of their rights, risks and implications.



---

## Mapping Contractual Agreements to Regulatory Documents

---

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 consider automated techniques to analyze personal customer data in order to achieve their business goals and unsurprisingly they do not always adhere to the law. In 2015, the Belgian privacy commission reported that Facebook's privacy policy breaches European law [118]. 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 and why [119]. As a result, people apply self-protection methods to ensure that their personal information is 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 Websites [120].

Chapter 4 and Chapter 5, focused on extracting valuable information from contractual agreements. The conducted experiments confirmed the subjective interpretation of legal texts and the complexity of reaching an agreement between different legal experts. In this Chapter we investigate the potential of mapping consent forms to regulatory documents to address research question 3:

### Research Question 3 (RQ3)

Given the subjective interpretability of legal texts, to what extent can we map contractual agreements to the applicable laws?

Once more, the running use-case for this Chapter will be privacy policies. Since privacy policies regulate the usage and maintenance of personal data, they must comply with the data protection laws. The supervisory authorities worldwide enact strict regulations regarding data protection of a natural person. The European Union, as well, has recently upgraded the current data protection directive to harmonize data privacy laws across Europe. The General Data Protection Regulation (GDPR) came into force since May 25th, 2018 [121]. All EU member states must now comply with GDPR and other corresponding regulatory documents. GDPR is a set of policies that guarantee users some rights regarding their personal data stored. Immediately after it came into force, groups of researchers from different domains started to analyze the various technical and legal challenges that this regulation would imply. In addition, many organizations and service providers had to apply major changes in their data processing and collection routines as well as modifying their human-readable privacy policy. Figure 6.1 shows GDPR attention in

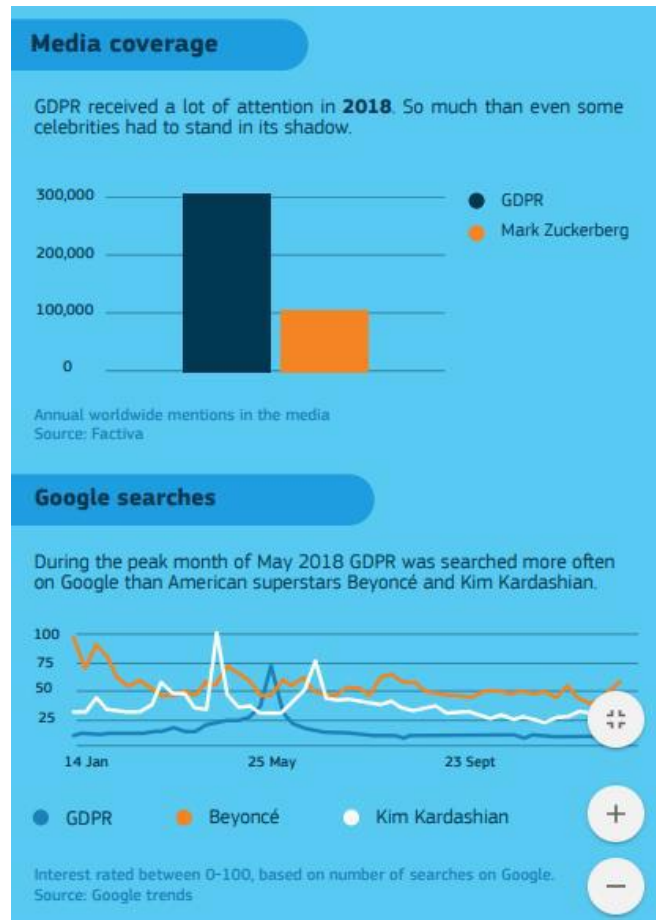


Figure 6.1: GDPR attention in 2018, published by the EC.

2018 in media coverage and Google searches.

In this Chapter we present an automatic technique for mapping privacy policies excerpts to relevant GDPR articles in order to assist users to familiarize themselves with their rights as a data subject and to support them in understanding their usage risk. *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. It takes a privacy policy and GDPR articles and finds the correlations between the two documents at the sentence and paragraph level. Since not all sentences in the privacy policy are related to data usage and analysis, first, the candidate sentences will be identified using a GATE pipeline [66]. This step will significantly reduce the number of processed excerpts in the policy text. After that, for each candidate, the most related Article (out of 99 GDPR Articles) 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.

This Chapter is based on the following publication [122] and is structured as follows: In Section 6.1, we describe *KnIGHT*'s workflow and its architecture. In Section 6.2 the conducted experiments are presented and finally Section 6.3 concludes this Chapter.

Table 6.1: Potential mappings between a privacy policy and the GDPR, based on our observation.

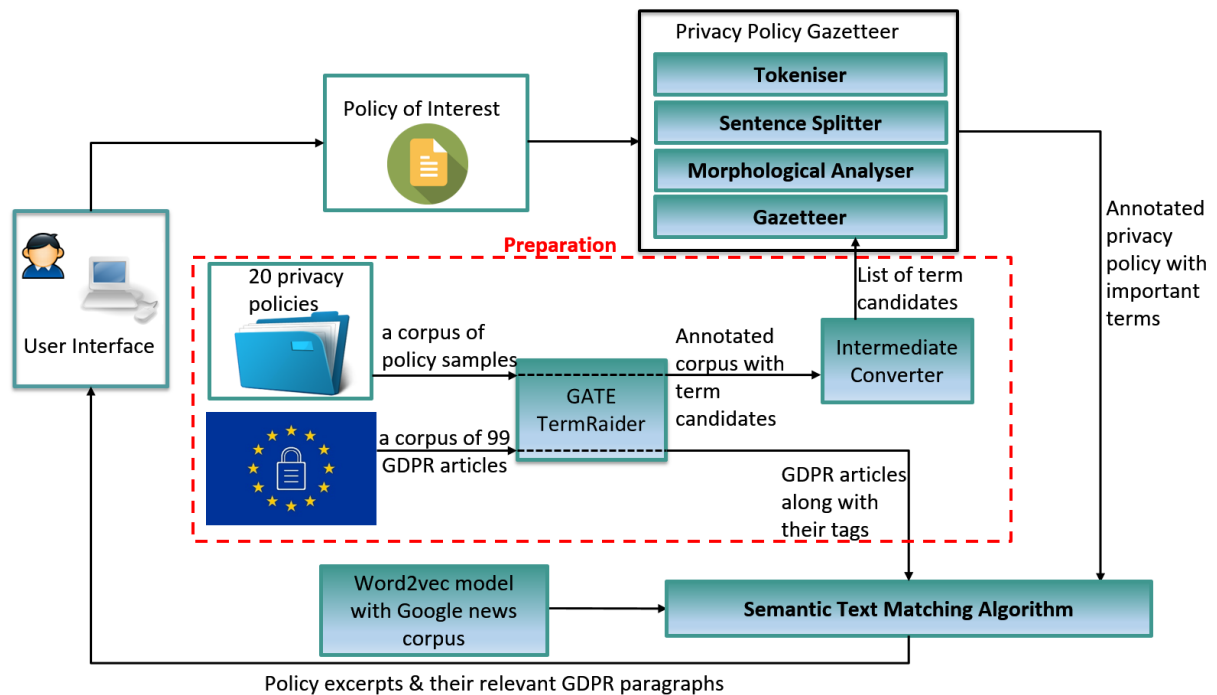
Privacy policy content	GDPR Description
Data category collected	Data subjects have the right to be informed about the collection and use of their personal data.
Goal of data collection	Organizations must provide transparent information about what they do with the personal data (right to be informed).
Third parties	The organization should provide data subjects the list of third parties that the personal data is shared with (right to be informed).
Retention period	Organizations must give details about how long they intend to keep the personal data.
Lawful basis for processing	Organizations must justify processing of personal data under six lawful bases (consent, contract, legal obligation, legitimate interest, vital interests, public task).
Security measures	Organizations must ensure that data is stored and processed safely & securely.
Automated decision making	Organizations must give details about how data is used for automatic analysis.
Complaint information	Organizations must clarify how users can make a complaint with a data protection authority.
Data access rights	Organizations are required to provide information about how data subject can see, change, delete, limit the use and export their personal data.

## 6.1 Mapping Privacy Policies to the GDPR

This section presents the architecture and implementation of *KnIGHT*. 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*. Table 6.1 shows some of the provisions that a privacy policy should contain according to the GDPR. This table inspired us to implement our initial idea of finding the relations between a privacy policy and the GDPR.

Figure 6.2 shows the architecture and workflow of *KnIGHT* which builds on GATE embedded and Deeplearning4j [123] 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<sup>1</sup>. The workflow consists of two main steps: the 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.

<sup>1</sup><https://code.google.com/archive/p/word2vec/>

Figure 6.2: Architecture and workflow of *KnIGHT*.

### 6.1.1 Pipeline 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*<sup>2</sup>. *TermRaider* is an English term extraction tool that runs over a text corpus and produces noun phrase term candidates together with a score that shows the salience of each term candidate in a domain-specific corpus. The preparation phase benefits from the *TermRaider* and include 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 that 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).

<sup>2</sup><https://gate.ac.uk/projects/neon/termraider.html>

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

### 6.1.2 Semantic Text Matching

Once the initial processing has been done, the system will be ready to accept the privacy policy. As mentioned before, *KnIGHT* relates a sentence in a policy to (a) paragraph(s) in GDPR. The rationale 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 more comfortable in any form of a page style. Furthermore, 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”.<sup>3</sup> Processing these kinds of sentences will only impose extra computation cost on the system without 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.

#### Privacy Policy Gazetteer

We have created a pipeline using GATE Embedded, which contains some basic pre-processing steps in NLP (tokeniser, sentence splitter, root finder) and a gazetteer that includes a list of essential terms in a privacy policy. As described in Section 6.1.1, 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: *Token* along with root feature; *Sentence*; and *Important Term*. If a sentence includes at least two important terms, it will be considered as a candidate.

#### 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 2 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  $T_{11}, \dots, T_{n1}$  and  $T_{12}, \dots, T_{m2}$ , the similarity between the two sets is calculated as shown in equation 6.1. In this formula,  $compositionalSim(Term1, Term2)$  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

<sup>3</sup><https://www.ryanair.com/gb/en/corporate/privacy-policy>

---

**Algorithm 2** Sketch of text matching algorithm.
 

---

**Require:** privacy policy candidate sentences, GDPR fingerprints, word2vec model

```

1: for all candidate sentences in the privacy policy do
2:   candidateSentence ← current sentence
3:   Set1 ← all important terms in candidateSentence
4:   MatchesList ← an empty list
5:   for all GDPR articles do
6:     ArticleNum ← article number
7:     Set2 ← article tags
8:     Sim ← similarity between Set1 & Set2
9:     if Sim > threshold then
10:      add Sim & ArticleNum to MatchesList
11:   SortedList ← Sort MatchesList acc. to sim
12:   BestArticleMatch ← SortedList[0]
13:   MaxSim ← 0
14:   Vec1 ← word2vec vector of candidateSentence
15:   for all paragraphs in BestArticleMatch do
16:     currPar ← current paragraph
17:     Vec2 ← word2vec vector of currPar
18:     Sim2 ← similarity between Vec1 & Vec2
19:     if Sim2 > MaxSim then
20:       MaxSim ← Sim2
21:       bestParMatch ← current paragraph

```

**Ensure:** Policy excerpts & their relevant GDPR paragraphs
 

---

(6.2) that 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, we apply the *Cosine* function again to calculate the similarity between two terms. Finally, if the similarity between two sets is greater than a fixed threshold (line 9 in algorithm 2), it will be added to a list along with the similarity score. 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.

$$Sim(S_1, S_2) = \frac{\sum_{i=1}^n \max_{1 \leq j \leq m} [CompositionalSim(T_{i1}, T_{j2})]}{\frac{n+m}{2}} \quad (6.1)$$

$$CompositionalSim(T_1, T_2) = cosineSim\left(\sum_{i=1}^n wordVector(T_{i1}), \sum_{j=1}^m wordVector(T_{j2})\right) \quad (6.2)$$

Having retrieved the best GDPR article for the current candidate sentence, the most related paragraph in the identified Article should be found (second step). Due to the lack of large domain-specific corpus, we have modified the word2vec model to be able to generate a vector for a sentence and paragraph. According to the literature, a simple yet efficient way to represent a sentence or a paragraph as a vector is computing the average of all word vectors [124]. 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 will lead to the identification of the paragraph with the highest similarity as the best match.

It is worth to mention that *KnIGHT* is policy and legislation independent. As an example, it can be applied to the cookie policy, which is sometimes embedded into the privacy policy itself. Cookies should comply with the ePrivacy directive [18] that will be soon replaced by proposed ePrivacy regulation [125]. In this case, a corpus of cookies policy in their natural text should be compiled to be ingested to *TermRaider*.

## 6.2 Experimental Study

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 several 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 four policies. That said, the primarily targeted end-users of the tool are non-experts. Therefore to contextualize these results, we conducted the second experiment: Two lay users were also asked to repeat the exercise. In this case, to have a realistic yardstick, they were instructed to spend between at least one and at most two hours to go through all four 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.

### 6.2.1 Posteriori Assessment

In order to perform a posteriori assessment, four 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. Successful execution of our pipeline generates some links between privacy policy sentences and GDPR paragraphs. Afterward, in order to reach a semi-conclusive result, four 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**<sup>4</sup>. 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 6.2 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

<sup>4</sup>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.

Table 6.2: 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	<b>6</b>	12	8	10	9	<b>9.75</b>	4	9	8	8	<b>7.25</b>
ResearchGate	29	11	14	8	14	<b>11.75</b>	8	9	12	4	<b>8.25</b>	10	6	9	11	<b>9</b>
Ryanair	10	2	3	2	3	<b>2.5</b>	4	4	4	3	<b>3.75</b>	4	3	4	4	<b>3.75</b>
Unilever	15	7	7	2	6	<b>5.5</b>	4	4	8	4	<b>5</b>	4	4	5	5	<b>4.5</b>

Table 6.3: 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
<b>Average</b>	<b>50</b>	<b>72.4</b>	<b>56.7</b>	<b>57.8</b>

Table 6.4: 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
<b>Average</b>	<b>33.4</b>	<b>67.7</b>	<b>46</b>	<b>46.4</b>

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 six annotations for **related** class in Avg column is the same annotations for all observers.

IAA is an agreement measure which can be calculated in Kappa or F-measure. When the observers have the choice to determine the span of the text for annotation, F-measure is recommended [95]. 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 [126]. The extension to more than two annotators is usually taken as the mean of the pair-wise agreements [127]. Furthermore, if the categories (A, B, C, ...) are ordered, weighted Kappa is considered [128]. Our three classes can be treated as an ordered list, because if one expert classifies a match into group **related** and the other into group **partially related**, this is closer than if one classifies into **related** and the other into **unrelated**.

Tables 6.3 and 6.4 show observed agreement and weighted kappa with linear weights. E1 to E4 represents 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 6.5. The first sentence from *Booking.com* informs the user that their personal data will only be

Table 6.5: Example of detected links &amp; experts (E1-E4) Assessments (R: related, P: partially related, U:unrelated).

Privacy Policy Sentence	Detected GDPR Paragraph	E1	E2	E3	E4
<i>Any additional personal details that you give us as a part of the market research will be used only with your consent.</i>	Article 7(3): <i>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.</i>	R	P	U	P
<i>We will comply with all applicable data protection laws and regulations and we will co-operate with data protection authorities.</i>	Article 77(1): <i>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.</i>	U	U	U	U
<i>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 where we feel it is appropriate to do so or where it is required by applicable laws and regulations.</i>	Article 8(1): <i>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 authorized 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.</i>	R	R	R	R

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 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 consent, the detected GDPR paragraph helps the end-user to be aware of their rights. Apart from the 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. Another issue identified was the generation of incomplete sets of tags for some GDPR articles. The second sentence in table 6.5 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. *KnIGHT* currently uses a general approach without any human involvement. Choosing 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 the detected paragraph is high, the degree of agreement increases. As an example, the third sentence in table 6.5 is the best match detected by *KnIGHT* with the similarity equals to 0.75 (max = 1) and it shows almost complete agreement.

Table 6.6: Average F-measure &amp; total time of 2 regular end-users annotations for 4 privacy policies.

	<b>Precision</b>	<b>Recall</b>	<b>F1</b>	<b>Time (min)</b>
<b>User1</b>	0.2	0.11	0.14	120
<b>User2</b>	0.46	0.08	0.14	30
<b><i>KnIGHT</i></b>	0.3	0.1	0.15	3

## 6.2.2 Potential End-Users Impact

According to the literature, end-users tend to skip privacy policies, and time plays a serious barrier in this case [129]. 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 Section 6.2.1 and GDPR. Here we have used the first expert (E1) annotations (in total 204 links) as a loose gold standard. Table 6.6 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 prove 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*.

The F-scores obtained in subsection 6.2.2 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.

## 6.3 Summary

In this Chapter, we presented *KnIGHT*, a tool for automatic mapping of privacy policies to GDPR. It is an initiative for privacy policies interpretation and helping regular end-users to familiarize themselves with respective data protection law in order to be aware of their rights as a citizen. Our 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.

Regarding the evaluation, four legal experts assessed and validated the automatic links, detected by

*KnIGHT*. The results showed 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. Furthermore, experimental results provide enough empirical evidence to answer RQ3 and showed that the interpretation of a 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 their rights as a data subject. Last but not least, we believe that the combination of *KnIGHT* and *Pripolis*, along with the assistance of legal experts, can be used as a recommender system to support regulatory experts in compliance checking.



## Conclusion

---

In this thesis, we considered the challenge of analysis and interpretation of contractual agreements, using knowledge extraction methods. Although regular end-users are the target category of this study, the research efforts described have implications for other kinds of users (e.g., lawyers, experts, compliance officers), since the described efforts can also be adapted for other use-cases with relatively minimal effort. After introducing the research problems, challenges and contributions in Chapter 1, providing necessary background concepts (cf. Chapter 2) and a literature survey (cf. Chapter 3), we addressed individual aspects of the study by proposing and evaluating novel solutions (or applications). In Chapter 4, we presented an ontology-based information extraction pipeline tailored specifically for end-user license agreements. Chapter 5 was dedicated to the analysis of privacy policies using ML methods, as one of the vital knowledge extraction methods supplementing natural language processing. Finally, given that all agreements between parties must comply with the relative legislation, Chapter 6 investigated the feasibility of mapping the human-readable agreements to the regulatory documents. In this Chapter, we conclude the thesis by reviewing once again the stated research questions, now taking into account the achieved individual results (cf. Section 7.1) and explore the opportunities for future work in Section 7.3.

### 7.1 Research Questions Analysis

To conclude this thesis, it is important to retrospectively revisit every research question, summarise the way they were tackled and interpret the results jointly. The main research problem was broken down into three more specific research questions (RQ1, RQ2, RQ3). The first research question (RQ1) aimed to identify efficient NLP techniques that are able to extract useful information for the benefit of regular end-users. The second research question (RQ2) explored whether the chosen approaches helps end-users to spend less time regarding contractual agreements comprehension. This research question explicitly addressed the practical aspect of our proposed solution. Finally, the third research question (RQ3) investigated a crucial aspect of every agreement, e.g., considering that all contracts and agreements should comply with the relevant regulations: to what extent can one create explicit mappings between contractual agreements and the applicable laws?

Research Question 1 (RQ1)

Are text mining techniques able to extract valuable information from contractual agreements?

In Chapters 4 and 5, we answered this question by exploring both rule-based and statistical NLP approaches. Chapter 4 presented an ontology-based annotation technique to extract deontic modalities (permission, prohibition, and duty) from the end-user license agreement. One of the major components of the ontology-based annotation pipeline is the ontology-aware gazetteer. This gazetteer is able to provide the ontological class of each entry and find mentions in the text, matching classes, instances, data property values, and labels in the ontology. Our ontology-based annotation pipeline relies on the ODRL ontology, a W3C recommendation that has gained huge endorsement in the semantic web community. Using GATE JAPE, a finite state transducer that operates over annotations based on regular expressions, we defined several linguistic rules according to the ODRL specification. Furthermore, the extracted deontic modalities were clustered based on their semantic similarity to shrink the amount of information presented to the users. On the other hand, Chapter 5 studied the first research question by scrutinizing machine learning and, more precisely, deep learning algorithms for classifying a privacy policy's paragraphs into pre-defined categories and attributes and estimating a policy's risk factor based on the predicted classes. Given that the BERT framework outperforms state-of-the-art results, we exploited the BERT's language model. The language model was pre-trained on large amounts of unlabeled data and is fine-tuned on specific labeled data to solve a downstream problem, which in our case is multi-label classification using the OPP-115 labeled dataset. Moreover, we used the predictions made by our trained models to assign five risk colors (Green, Yellow, Red) to the five privacy icons (*Expected Use*, *Expected Collection*, *Precise Location*, *Data Retention* and *Children Privacy*). Our risk color assignment rules follow a baseline definition, founded on the *Disconnect* privacy icons, and outperforms the baseline's assignment accuracy. The challenges and difficulties in reproducing previous results and creating different gold standards, motivated us to establish a strong baseline for privacy policy classification and analysis in Section 5.1. We have made our implementation and resources open and accessible to the relevant community. In conclusion, the experimental studies presented in both chapters indicate that in the presence of a reliable dataset, supervised machine learning is favored over the hand-coded rules. However, due to lack of such reliable gold standard in several areas of legal domain, and given the clear terminology of legal documents, defining hand-coded linguistic rules are a suitable replacement and ensure promising results.

Research Question 2 (RQ2)

Does ontology-based information extraction help end-users to spend less time to understand contractual agreements?

Chapter 4 addressed this question by presenting quantitative and qualitative evaluation to measure the practical functionality of our proposed solution. In particular, the experiments targeted the OBIE approach performance. However, we believe that they provide enough evidence for addressing the second research question. The experiments were designed to estimate whether our approach helps users to understand license agreements. Although due to funding restrictions, the number of selected agreements and participants was the bare minimum required for an experiment of this kind, the results are still



able to indicate the value in extending and improving our approach. The experiments indicate that our initial hypotheses holds water, i.e., even though semi-automatic approaches are effected by information loss, they considerably save time and effort spent by users to arrive at a similar level of understanding. Given that a majority of end-users ignore reading consent forms and agreements, our proposed approach motivates them to quickly familiarize themselves with the terms and conditions they are agreeing to.

### Research Question 3 (RQ3)

Given the subjective interpretability of legal texts, to what extent can we map contractual agreements to the applicable laws?

This research question was investigated in Chapter 6. As state-of-the-art approaches have not prioritized the regular end-users needs and concerns, we designed a policy reading system to assist consumers in familiarizing themselves with a specific policy and the relevant regulation. Our goal was 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. Although our proposed approach is general, we specifically focused on privacy policies for two main reasons: first, since privacy policies regulate the use of personal data and information, they have to comply with a small set of regulations (as opposed to EULAs); and second, due to the high impact of the GDPR on data and privacy agreements, many services have updated the human-readable privacy policies and changed their specifics according to the GDPR. Hence, mapping privacy policies to the GDPR is a concrete and current use-case for our problem. Using semantic text matching, we introduced a pipeline that first finds the candidate sentences which have the potential to be mapped to the GDPR articles. Then a scoring algorithm computes the text similarity between the candidate sentences and the relevant GDPR paragraphs. The experiments conducted, confirm the complexity of the task and the subjectivity of human judgment. Somewhat counter-intuitively, we observe that agreement between the experts is generally harder to achieve than between average users. This is probably due to the experts' higher understanding and the ability for more critical inspection of legal texts. The discrepancies between different expert's judgments proved the high subjectivity of such interpretation. According to the experts, the drafting of mappings between a contractual agreement and the regulatory documents strongly depends on what one considers as being 'translated into a contractual agreement'. In some cases, the content of an agreement will duly reflect the wording of the corresponding regulation's provision. In others, the link between the legal text and the contractual agreement might not appear as striking. As mentioned earlier, our approach relies on semantic similarity between words and sentences to actually match excerpts from privacy policies with their legislative counterparts. However, the evaluation proved that a mere text matching between the privacy policy and the GDPR is likely to prove incomplete if the semantic of these provisions is left aside. We learned from the experts' annotations that while some provisions from the GDPR are not meant to be translated into a contractual agreement between the controller and the data subject at all, others might be transposed throughout the entire privacy policy without, however, being equivocally phrased. For instance, in the case of the GDPR, according to the *purpose limitation principle* (art. 5(1)b) and the *lawfulness of processing* (art. 6), personal data may only be processed on the basis of a valid legal ground, and for (a) specific purpose(s). Yet, most controllers spread that information across their entire privacy notice without centralizing this information into a single, comprehensive chart/table. Hence, during our experiments, we noticed that there are many implicit matching in the experts' annotations to these articles that our tool failed to detect. Last but not least, the normative text is often complemented by

**soft-law instruments.** These opinions and guidelines substantiate and particularise the somehow vague and imprecise rules laid down in legislative materials. Therefore, while it is not always feasible to link excerpts of contractual agreements directly with provisions from regulations, it might be practical to look for a relevant match in the soft-law instruments.

## 7.2 Limitations

Research efforts that could build on or extend the reported results should consider three major limitations that were identified throughout the course of our research. The first limitation relates to the incomplete set of linguistic rules in *EULAide* as well as the limited number of instances and concepts in ODRL ontology. Given that the ontology-aware gazetteer annotates the license agreement with the ontology instances, the pipeline fails to notice a deontic modality when an instance is not present in the ontology. Moreover, without consulting the domain experts it is not feasible to add the missing instances to the ontology. A constant non-technical challenge in legal text interpretation is to attain commitment from domain experts on a voluntary basis.

The second limitation relates to the limited number of participants in our usability experiments. Contractual agreements are long, complex and exhibit legal lingua. Therefore, it was challenging to find volunteers who are willing to take part in our experiments *pro bono*. With sufficient funding, a more comprehensive set of experiments designed around our own method can be carried out.

The third limitation is based on the fact that textual descriptions of rules and regulations remain subject to interpretation, which makes our mapping approach to align excerpts in contractual agreements to sections in regulatory documents difficult and incomplete. A possible solution is to combine text matching algorithms with ontologies and vocabularies that encode the semantic of provisions and regulations. In the case of GDPR, there have been initial efforts to represent its article in the form of vocabularies and ontologies [130, 131]. Exploiting the domain specific ontologies will enable deeper analysis of the texts and extract more knowledge from respective regulations.

## 7.3 Closing Remarks and Future Work

This thesis considers whether automated methods can be tailored to target difficult-to-read contractual agreements and present them in a more user-friendly manner to minimise risks and legal pitfalls that can ensue without proper interpretation. Despite the challenges and difficulties, our results indicate that NLP techniques combined with OBIE and ML can be beneficial to support legal text comprehension.

In this last section, we discuss possible improvements, extensions, and long term opportunities that open up after our work. Based on our experience, funding opportunities should be sought to improve the broadness of the presented research efforts, after the application and evaluation in a limited number of specific use-cases was found to be of benefit to the targeted end-users.

1. *Future of regulatory documents:* currently, the regulatory documents enforce specific regulations, and every agreement and contract between parties must adhere to those documents. However, there are no clear guidelines on the form and contents of policies and agreements that need to comply with the regulations. This is basically the main reason that there exist various templates and formats for any kind of agreement. Such guidelines can significantly enhance levels of readability and compliance of agreements and diminish their ambiguity. In addition, enforcing parties to publish a machine-processable agreement or policies along with a human-readable version, could open a

new path for automatic analysis and interpretation of those documents and could lead to practical scanning tools for supervisory authorities.

2. *Future of contracts and agreements*: as studied in this thesis, there are a number of established standards for defining licenses and agreements. Moreover, in the past few years, several tools and frameworks were implemented based on these standards to assist companies and service providers produce a machine-readable agreement. Unsurprisingly, the majority of enterprises only provide a human-readable agreement. The fact that machine-readable agreements could envision the creation of apps and browser extensions for further analysis are regularly underestimated. Such apps and plugins enable end-users to quickly check what they are agreeing to, instead of reading the whole text. Service providers are able to create a machine-readable agreement with no big effort.
3. *Future of law automation*: the experiments conducted in this thesis indicated that (semi-)automatic analysis of contractual agreements gains promising results. Although in some areas, more data is required in order to obtain higher quality results. High-quality results could promote recommender systems for supervisory authorities to monitor the agreements' compliance and could significantly reduce the required human effort. Such automatic methods could automatically notify supervisory authorities about potentially unlawful content or data processing activities. The organizations can also control whether the actual services provided to the consumers are in line with the human-readable policies. Last but not least, providing diverse vocabulary-based tools and frameworks (e.g., license generators, privacy policy generator) motivate companies to publish a machine-readable agreement along with the human-readable text.



# Bibliography

---

- [1] *General Data Protection Regulation*, URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593420699390&uri=CELEX:32016R0679> (cit. on pp. 1, 12).
- [2] J. A. Obar and A. Oeldorf-Hirsch, *The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services*, Information, Communication & Society (2018) 1 (cit. on pp. 1, 63).
- [3] *Skandia Takes the Terminal out of Terms and Conditions*, 2011, URL: <https://www.prnewswire.co.uk/news-releases/skandia-takes-the-terminal-out-of-terms-and-conditions-145280565.html> (visited on 23/05/2011) (cit. on p. 1).
- [4] U. Benoliel and S. I. Becher, *The Duty to Read the Unreadable*, Boston College law review (2019) (cit. on p. 3).
- [5] A. M. McDonald and L. F. Cranor, *The cost of reading privacy policies*, ISJLP 4 (2008) 543 (cit. on pp. 3, 63).
- [6] M. ISO, *Ergonomics of human-system interaction—part 11: Usability: Definitions and concepts*, 2018 (cit. on p. 3).
- [7] J. M. Balkin, *Understanding Legal Understanding: The Legal Subject and the Problem of Legal Coherence*, The Yale Law Journal 103 (1993) 105 (cit. on p. 3).
- [8] L. Rodak, “Objective Interpretation as Conforming Interpretation”, *Oñati Socio-Legal Series*, 2012 (cit. on p. 4).
- [9] Y. Zhang and B. C. Wallace, *A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification*, CoRR abs/1510.03820 (2015), URL: <http://arxiv.org/abs/1510.03820> (cit. on pp. 4, 22, 23).
- [10] Y. Kim, “Convolutional Neural Networks for Sentence Classification”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014 1746 (cit. on pp. 4, 64).
- [11] T. Young, D. Hazarika, S. Poria and E. Cambria, *Recent Trends in Deep Learning Based Natural Language Processing [Review Article]*, IEEE Comput. Intell. Mag. 13 (2018) 55 (cit. on p. 4).
- [12] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai and T. Chen, *Recent advances in convolutional neural networks*, Pattern Recognit. 77 (2018) 354 (cit. on p. 4).

- [13] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019 4171 (cit. on pp. 4, 59, 63, 65, 66).
- [14] J. A. Obar and A. Oeldorf-Hirsch, *The Clickwrap: A Political Economic Mechanism for Manufacturing Consent on Social Media*, *Social Media + Society* **4** (2018) 1 (cit. on p. 11).
- [15] J. C. Gooch, *Bell, T. W. (2014). Intellectual Privilege: Copyright, Common Law, and the Common Good*, *Journal of Technical Writing and Communication* **45** (2015) 323 (cit. on p. 11).
- [16] *Computer Programs Directive*, URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1589549912653&uri=CELEX:32009L0024> (cit. on p. 11).
- [17] A. A. of Law Libraries, A. L. Association, A. of Academic Health Sciences Libraries, A. of Research Libraries and S. L. Association, *Principles for Licensing Electronic Resources: Final Draft*, Association of Research Libraries, 1997 (cit. on p. 12).
- [18] *Privacy and Electronic Communications Directive*, URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1539354724573&uri=CELEX:32002L0058> (cit. on pp. 12, 81).
- [19] J. K. Sørensen and S. Kosta, “Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites”, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 2019 1590 (cit. on pp. 12, 66).
- [20] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub and T. Holz, *We Value Your Privacy ... Now Take Some Cookies - Measuring the GDPR's Impact on Web Privacy*, *Informatik Spektrum* **42** (2019) 345 (cit. on pp. 12, 66).
- [21] T. Linden, H. Harkous and K. Fawaz, *The Privacy Policy Landscape After the GDPR*, *CoRR* **abs/1809.08396** (2018), URL: <http://arxiv.org/abs/1809.08396> (cit. on pp. 12, 36, 66).
- [22] OECD, *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, 2002 (cit. on p. 12).
- [23] S. Sandeen, *The Sense and Nonsense of Web Site Terms of Use Agreements*, *Hamline Law Review* **26** (2003) 499 (cit. on p. 13).
- [24] *RDF Schema 1.1 Recommendation*, URL: <https://www.w3.org/TR/2014/REC-rdf-schema-20140225/> (cit. on pp. 13, 14).
- [25] *Extensible Markup Language (XML) 1.0 (Fifth Edition)*, URL: <https://www.w3.org/TR/2008/REC-xml-20081126/> (cit. on p. 14).

- 
- [26] *ECMA-404: The JSON Data Interchange Format*, URL: <http://www.ecma-international.org/publications/standards/Ecma-404.htm> (cit. on p. 14).
- [27] T. R. Gruber, *A Translation Approach to Portable Ontologies*, *Knowledge Acquisition* **5** (1993) (cit. on p. 15).
- [28] M. L. H. Abelson B. Adida and N. Yergler, *ccREL: The creative commons rights expression language. Technical report*, tech. rep., Creative Commons, 2008 (cit. on pp. 16, 41).
- [29] *ODRL Version 2.2 Ontology*, 2017, URL: <https://www.w3.org/ns/odrl/2/> (visited on 30/06/2020) (cit. on pp. 17, 41).
- [30] *ISO/IEC 21000-6 Rights Data Dictionary*, 2005, URL: <http://iso21000-6.net/view/rddDictionary.php> (visited on 30/06/2020) (cit. on pp. 17, 41).
- [31] J. Polo, J. Prados and J. Delgado, “Interoperability between ODRL and MPEG-21 REL”, *Proceedings of the First International Workshop on the Open Digital Rights Language (ODRL), Vienna, Austria, April 22-23, 2004*, 2004 65 (cit. on p. 18).
- [32] *W3C, The Platform for Privacy Preferences (P3P), W3C Recommendation*, URL: <http://www.w3.org/TR/P3P> (cit. on p. 19).
- [33] *Enterprise Privacy Authorization Language (EPAL 1.2)*, URL: <https://www.w3.org/Submission/2003/SUBM-EPAL-20031110/> (cit. on p. 19).
- [34] D. Britz, *Understanding Convolutional Neural Networks for NLP*, 2015, URL: <http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/> (visited on 30/06/2020) (cit. on p. 24).
- [35] M. Palmirani, T. Ognibene and L. Cervone, “Legal Rules, Text and Ontologies Over Time”, *Proceedings of the RuleML2012@ECAI Challenge, at the 6th International Symposium on Rules, Montpellier, France, August 27th-29th, 2012*, 2012 (cit. on p. 25).
- [36] J. Breuker, A. Boer, R. Hoekstra and K. van den Berg, “Developing Content for LKIF: Ontologies and Frameworks for Legal Reasoning”, *Legal Knowledge and Information Systems - JURIX 2006: The Nineteenth Annual Conference on Legal Knowledge and Information Systems, Paris, France, 7-9 December 2006*, 2006 169 (cit. on p. 25).
- [37] R. Hoekstra, J. Breuker, M. D. Bello and A. Boer, “The LKIF Core Ontology of Basic Legal Concepts”, *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques June 4th, 2007, Stanford University, Stanford, CA, USA*, 2007 43 (cit. on pp. 25, 32).
- [38] T. M. van Engers, A. Boer, J. Breuker, A. Valente and R. Winkels, “Ontologies in the Legal Domain”, *Digital Government: E-Government Research, Case Studies, and Implementation*, 2008 233 (cit. on p. 25).
- [39] G. Sartor, *Legal concepts as inferential nodes and ontological categories*, *Artif. Intell. Law* **17** (2009) 217 (cit. on p. 25).
- [40] *Semantics Of Business Vocabulary And Rules*, URL: <https://www.omg.org/spec/SBVR/About-SBVR/> (cit. on p. 27).

- [41] E. Abi-Lahoud, T. Butler, D. Chapin and J. Hall, “Interpreting Regulations with SBVR”, *Joint Proceedings of the 7th International Rule Challenge, the Special Track on Human Language Technology and the 3rd RuleML Doctoral Consortium, Seattle, USA, July 11 -13, 2013*, 2013 (cit. on p. 28).
- [42] *SPARQL 1.1 Query Language*, URL: <https://www.w3.org/TR/sparql11-query/> (cit. on p. 29).
- [43] T. Butler, E. Abi-Lahoud and A. Espinoza, “Designing Semantic Technologies for Regulatory Change Management in the Financial Industry”, *ICIS*, 2015 (cit. on p. 29).
- [44] A. van Lamsweerde, “Goal-Oriented Requirements Engineering: A Guided Tour”, *5th IEEE International Symposium on Requirements Engineering (RE 2001), 27-31 August 2001, Toronto, Canada*, 2001 249 (cit. on p. 29).
- [45] A. I. Antón, J. B. Earp, D. Bolchini, Q. He, C. Jensen, W. Stufflebeam and T. N. F. Standardization, *The lack of clarity in financial privacy policies and the need for standardization*, tech. rep., IEEE Security and Privacy, 2003 (cit. on p. 29).
- [46] T. D. Breaux and A. I. Antón, “Deriving Semantic Models from Privacy Policies”, *6th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY 2005), 6-8 June 2005, Stockholm, Sweden*, 2005 67 (cit. on p. 30).
- [47] T. D. Breaux and A. I. Antón, “Analyzing Goal Semantics for Rights, Permissions, and Obligations”, *13th IEEE International Conference on Requirements Engineering (RE 2005), 29 August - 2 September 2005, Paris, France*, 2005 177 (cit. on p. 30).
- [48] T. D. Breaux and A. I. Antón, *Analyzing Regulatory Rules for Privacy and Security Requirements*, IEEE Trans. Software Eng. **34** (2008) 5 (cit. on p. 30).
- [49] K. Sapkota, A. Aldea, M. Younas, D. A. Duce and R. Bañares-Alcántara, “Extracting meaningful entities from regulatory text: Towards automating regulatory compliance”, *Fifth IEEE International Workshop on Requirements Engineering and Law, RELAW 2012, Chicago, IL, USA, September 25, 2012*, 2012 29 (cit. on p. 31).
- [50] M. B. Sesen, P. Suresh, R. Bañares-Alcántara and V. Venkatasubramanian, *An ontological framework for automated regulatory compliance in pharmaceutical manufacturing*, Comput. Chem. Eng. **34** (2010) 1155 (cit. on p. 32).
- [51] N. Zeni, N. Kiyavitskaya, L. Mich, J. R. Cordy and J. Mylopoulos, *GaiusT: supporting the extraction of rights and obligations for regulatory compliance*, Requir. Eng. **20** (2015) 1 (cit. on p. 32).
- [52] N. Kiyavitskaya, N. Zeni, J. R. Cordy, L. Mich and J. Mylopoulos, *Cerno: Light-weight tool support for semantic annotation of textual documents*, Data Knowl. Eng. **68** (2009) 1470 (cit. on p. 32).
- [53] T. D. Breaux, M. W. Vail and A. I. Antón, “Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations”, *14th IEEE International Conference on Requirements Engineering (RE 2006), 11-15 September 2006, Minneapolis/St.Paul, Minnesota, USA*, 2006 46 (cit. on p. 32).



- 
- [54] K. D. Ashley,  
*Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*,  
Cambridge University Press, 2017 (cit. on p. 33).
- [55] M. Kim and R. Goebel,  
“Two-step cascaded textual entailment for legal bar exam question answering”,  
*Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*, 2017 283 (cit. on p. 33).
- [56] I. Chalkidis, I. Androutopoulos and A. Michos, “Extracting contract elements”,  
*Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*, 2017 19 (cit. on p. 33).
- [57] W. Alschner and D. Skougarevskiy,  
“Towards an automated production of legal texts using recurrent neural networks”,  
*Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*, 2017 229 (cit. on p. 33).
- [58] E. Cabrio, A. P. Apro시오 and S. Villata, “These Are Your Rights - A Natural Language Processing Approach to Automated RDF Licenses Generation”,  
*The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, 2014 255 (cit. on p. 33).
- [59] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. H. Hovy, J. R. Reidenberg and N. M. Sadeh,  
“The Creation and Analysis of a Website Privacy Policy Corpus”,  
*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016 1330 (cit. on pp. 33, 60, 62, 63, 68, 71).
- [60] N. Guntamukkala, R. Dara and G. Gréwal, “A Machine-Learning Based Approach for Measuring the Completeness of Online Privacy Policies”,  
*14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, 2015 289 (cit. on p. 33).
- [61] E. Costante, Y. Sun, M. Petković and J. den Hartog,  
“A Machine Learning Solution to Assess Privacy Policy Completeness: (Short Paper)”,  
*Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, 2012 91 (cit. on p. 33).
- [62] R. N. Zaeem, R. L. German and K. S. Barber,  
*PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining*,  
*ACM Trans. Internet Techn.* **18** (2018) 53:1 (cit. on pp. 34, 66).
- [63] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto and J. Serna, “PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation”,  
*Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, IWSPA@CODASPY 2018, Tempe, AZ, USA, March 19-21, 2018*, 2018 15 (cit. on pp. 34, 66).
- [64] I. Chalkidis, I. Androutopoulos and A. Michos,  
*Obligation and Prohibition Extraction Using Hierarchical RNNs*, *CoRR* **abs/1805.03871** (2018) (cit. on p. 35).

- [65] J. O’Neill, P. Buitelaar, C. Robin and L. O’Brien, “Classifying sentential modality in legal language: a use case in financial regulations, acts and directives”, *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017*, 2017 159 (cit. on p. 35).
- [66] H. Cunningham, D. Maynard and K. Bontcheva, *Text Processing with GATE (Version 8)*, Gateway Press CA, 2011 (cit. on pp. 35, 40, 50, 76).
- [67] H. Harkous, K. Fawaz, R. Lebre, F. Schaub, K. G. Shin and K. Aberer, “Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning”, *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, 2018 531 (cit. on pp. 36, 60, 63, 64, 66, 67).
- [68] S. Alzahrani, N. Salim and A. Abraham, *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods*, *IEEE Trans. Syst. Man Cybern. Part C* **42** (2012) 133 (cit. on p. 37).
- [69] S. Sunkle, D. Kholkar and V. Kulkarni, “Toward Better Mapping between Regulations and Operational Details of Enterprises Using Vocabularies and Semantic Similarity”, *Proceedings of the CAiSE 2015 Forum at the 27th International Conference on Advanced Information Systems Engineering co-located with 27th International Conference on Advanced Information Systems Engineering (CAiSE 2015), Stockholm, Sweden, June 10th, 2015*, 2015 229 (cit. on p. 37).
- [70] C. Erl, *Semantic Text Matching of Company Policies and Regulatory Documents using Text Similarity Measures*, 2018,  
URL: <https://www.matthes.in.tum.de/pages/1nbhr3lcnxo4b/Master-s-Thesis-von-Christoph-Erl> (cit. on p. 37).
- [71] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H. Micklitz, G. Sartor and P. Torroni, *CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service*, *Artif. Intell. Law* **27** (2019) 117 (cit. on p. 37).
- [72] F. Lagioia, F. Ruggeri, K. Drazewski, M. Lippi, H. Micklitz, P. Torroni and G. Sartor, “Deep Learning for Detecting and Explaining Unfairness in Consumer Contracts”, *Legal Knowledge and Information Systems - JURIX 2019: The Thirty-second Annual Conference, Madrid, Spain, December 11-13, 2019*, 2019 43 (cit. on p. 37).
- [73] G. Contissa, K. Docter, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Pałka, G. Sartor and P. Torroni, *Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence*, Available at SSRN 3208596 (2018) (cit. on pp. 38, 66).
- [74] N. M. Nejad, S. Scerri, S. Auer and E. M. Sibarani, “EULAide: Interpretation of End-User License Agreements using Ontology-Based Information Extraction”, *Proceedings of the 12th International Conference on Semantic Systems, SEMANTICS 2016, Leipzig, Germany, September 12-15, 2016*, 2016 73 (cit. on p. 40).
- [75] N. M. Nejad, S. Scerri and S. Auer, “Semantic Similarity based Clustering of License Excerpts for Improved End-User Interpretation”, *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, 2017 144 (cit. on p. 40).

- 
- [76] S. Steyskal and A. Polleres, “Defining expressive access policies for linked data using the ODRL ontology 2.0”, *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, 2014 20 (cit. on p. 40).
- [77] P. A. Jamkhedkar and G. L. Heileman, “A formal conceptual model for rights”, *Proceedings of the 8th ACM Workshop on Digital Rights Management, Alexandria, VA, USA, October 27, 2008*, 2008 29 (cit. on p. 40).
- [78] E. Daga, M. d’Aquin, E. Motta and A. Gangemi, “A Bottom-Up Approach for Licences Classification and Selection”, *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers*, 2015 257 (cit. on p. 40).
- [79] *Linked Data Rights*, 2014,  
URL: <http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/> (visited on 30/06/2020) (cit. on p. 41).
- [80] F. G. Serena Villata, *L4LOD Vocabulary Specification 0.2*, 2013,  
URL: [http://ns.inria.fr/l4lod/v2/l4lod\\_v2.html](http://ns.inria.fr/l4lod/v2/l4lod_v2.html) (visited on 30/06/2020) (cit. on p. 41).
- [81] L. Dodds, *Open Data Rights Statement Vocabulary*, 2013,  
URL: <http://schema.theodi.org/odrs/> (visited on 30/06/2020) (cit. on p. 41).
- [82] *Copyright Ontology*, 2019,  
URL: <http://rhizomik.net/html/ontologies/copyrightonto/> (visited on 30/06/2020) (cit. on p. 41).
- [83] R. G. Jaime Delgado, *Intellectual Property Rights Ontology (IPRonto)*, 2013,  
URL: <http://dmag.ac.upc.edu/ontologies/ipronto/> (visited on 30/06/2020) (cit. on p. 41).
- [84] D. M. H. Cunningham and V. Tablan, *JAPE: a Java Annotation Patterns Engine (Second Edition)*, tech. rep., Department of Computer Science, University of Sheffield, 2000 (cit. on p. 40).
- [85] P. Kolb, *Disco: A multilingual database of distributionally similar words*, *Proceedings of KONVENS-2008, Berlin (2008)* (cit. on p. 45).
- [86] A. Qadir, P. N. Mendes, D. Gruhl and N. Lewis, “Semantic Lexicon Induction from Twitter with Pattern Relatedness and Flexible Term Length”, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, 2015 2432 (cit. on p. 45).
- [87] N. N. Chan, A. Roussanally and A. Boyer, “Learning Resource Recommendation: An Orchestration of Content-Based Filtering, Word Semantic Similarity and Page Ranking”, *Open Learning and Teaching in Educational Communities - 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Graz, Austria, September 16-19, 2014, Proceedings*, 2014 302 (cit. on p. 45).
- [88] S. Bhagwani, S. Satapathy and H. Karnick, “sranjans : Semantic Textual Similarity using Maximal Weighted Bipartite Graph Matching”, *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, 2012 579 (cit. on p. 45).

- [89] N. K. Nagwani and S. Verma, *A frequent term and semantic similarity based single document text summarization algorithm*, International Journal of Computer Applications (0975–8887) Volume (2011) 36 (cit. on p. 45).
- [90] N. Lavesson, M. Boldt, P. Davidsson and A. Jacobsson, *Learning to detect spyware using end user license agreements*, Knowl. Inf. Syst. **26** (2011) 285 (cit. on p. 45).
- [91] C. C. Aggarwal and C. Zhai, “A Survey of Text Clustering Algorithms”, *Mining Text Data*, Springer, 2012 77 (cit. on pp. 46, 52).
- [92] Y. Zhao, G. Karypis and U. M. Fayyad, *Hierarchical Clustering Algorithms for Document Datasets*, Data Min. Knowl. Discov. **10** (2005) 141 (cit. on p. 46).
- [93] A. K. A. Mohamed, *Optimizing the Usability of an EULA Interpretation Service on Mobile Devices*, master thesis: University of Bonn, 2017 (cit. on p. 47).
- [94] *A framework for building native apps using React*, 2015, URL: <https://github.com/facebook/react-native> (visited on 30/06/2020) (cit. on p. 47).
- [95] G. Hripcsak and A. S. Rothschild, *Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval*, JAMIA **12** (2005) 296 (cit. on pp. 50, 82).
- [96] H. Cunningham, “Information Extraction, Automatic”, *Encyclopedia of Language and Linguistics*, Elsevier, 2006 665 (cit. on p. 50).
- [97] T. Kenter and M. de Rijke, “Short Text Similarity with Word Embeddings.”, *CIKM*, 2015 1411 (cit. on p. 52).
- [98] G. Tsatsaronis, I. Varlamis and M. Vazirgiannis, *Text Relatedness Based on a Word Thesaurus*, CoRR **abs/1401.5699** (2014) (cit. on p. 52).
- [99] A. Khan, N. Salim and Y. J. Kumar, *A framework for multi-document abstractive summarization based on semantic role labelling*, Appl. Soft Comput. **30** (2015) 737 (cit. on p. 52).
- [100] R. M. Aliguliyev, *A new sentence similarity measure and sentence based extractive technique for automatic text summarization*, Expert Syst. Appl. **36** (2009) 7764 (cit. on p. 52).
- [101] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to information retrieval*, Cambridge University Press, 2008 (cit. on pp. 53, 68).
- [102] N. M. Nejad, P. Jabat, R. Nedelchev, S. Scerri and D. Graux, “Establishing a Strong Baseline for Privacy Policy Classification”, *ICT Systems Security and Privacy Protection*, 2020 370 (cit. on p. 59).
- [103] N. M. Nejad, D. Graux and D. Collarana, “Towards Measuring Risk Factors in Privacy Policies”, *Proceedings of the Workshop on Artificial Intelligence and the Administrative State co-located with 17th International Conference on AI and Law (ICAAIL 2019), Montreal, QC, Canada, June 17, 2019*, 2019 18 (cit. on p. 59).
- [104] I. Pollach, *What’s wrong with online privacy policies?*, Commun. ACM **50** (2007) 103 (cit. on p. 63).

- 
- [105] A. Mnih and G. Hinton, “Three New Graphical Models for Statistical Language Modelling”, *Proceedings of the 24th International Conference on Machine Learning*, 2007 641 (cit. on p. 63).
- [106] R. Collobert and J. Weston, “A unified architecture for natural language processing: deep neural networks with multitask learning”, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, 2008 160 (cit. on p. 63).
- [107] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2013 3111 (cit. on p. 63).
- [108] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, “Bag of Tricks for Efficient Text Classification”, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, 2017 427 (cit. on p. 63).
- [109] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, *Enriching Word Vectors with Subword Information*, *TACL* **5** (2017) 135 (cit. on pp. 63, 64).
- [110] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu and B. Qin, “Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, 2014 1555 (cit. on p. 64).
- [111] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015 (cit. on p. 64).
- [112] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention is All you Need”, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017 5998 (cit. on pp. 65, 66).
- [113] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., *Google’s neural machine translation system: Bridging the gap between human and machine translation*, *CoRR* **abs/1609.08144** (2016) (cit. on p. 65).
- [114] F. Sebastiani, *Machine learning in automated text categorization*, *ACM Comput. Surv.* **34** (2002) 1 (cit. on p. 68).
- [115] E. Wiener, J. O. Pedersen, A. S. Weigend et al., “A neural network approach to topic spotting”, *Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval*, 1995 332 (cit. on p. 68).
- [116] V. Van Asch, *Macro-and micro-averaged evaluation measures [[basic draft]]*, Belgium: CLiPS (2013) (cit. on p. 68).
- [117] Y. You, J. Li, J. Hseu, X. Song, J. Demmel and C. Hsieh, *Reducing BERT Pre-Training Time from 3 Days to 76 Minutes*, *CoRR* **abs/1904.00962** (2019) (cit. on p. 68).



- [118] *Facebook's privacy policy breaches European law, report finds*, 2015, URL: <https://www.theguardian.com/technology/2015/feb/23/facebooks-privacy-policy-breaches-european-law-report-finds> (visited on 30/06/2020) (cit. on p. 75).
- [119] D. Meyer, *Microsoft's Windows 10 breaches data protection law, say Dutch regulator*, 2017, URL: <https://www.zdnet.com/article/microsofts-windows-10-breaches-data-protection-law-say-dutch-regulator/> (visited on 30/06/2020) (cit. on p. 75).
- [120] A. Acquisti and J. Grossklags, *Privacy and Rationality in Individual Decision Making*, *IEEE Security and Privacy* **3** (2005) 26 (cit. on p. 75).
- [121] P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR)*, A Practical Guide, 1st Ed., Cham: Springer International Publishing (2017) (cit. on p. 75).
- [122] N. M. Nejad, S. Scerri and J. Lehmann, "KnIGHT: Mapping Privacy Policies to GDPR", *Knowledge Engineering and Knowledge Management - 21st International Conference, EKAW 2018, Nancy, France, November 12-16, 2018, Proceedings*, 2018 258 (cit. on p. 76).
- [123] *DeepLearning4j: Open-source distributed deep learning for the JVM Apache Software Foundation License 2.0*, 2015, URL: <http://deeplearning4j.org/word2vec.html> (cit. on p. 77).
- [124] T. Kenter, A. Borisov and M. de Rijke, "Siamese CBOW: Optimizing Word Embeddings for Sentence Representations", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016 941 (cit. on p. 80).
- [125] *Regulation on Privacy and Electronic Communications*, URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:0010:FIN> (cit. on p. 81).
- [126] G. Hripcsak and D. F. Heitjan, *Measuring agreement in medical informatics reliability studies*, *Journal of Biomedical Informatics* **35** (2002) 99 (cit. on p. 82).
- [127] J. L. Fleiss, *Measuring Agreement between Two Judges on the Presence or Absence of a Trait*, *Biometrics* **31** (1975) 651 (cit. on p. 82).
- [128] J. M. Cohen, *Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit.*, *Psychological bulletin* **70** **4** (1968) 213 (cit. on p. 82).
- [129] L. F. Cranor, M. Arjula and P. Guduru, "Use of a P3P user agent by early adopters", *Proceedings of the 2002 ACM Workshop on Privacy in the Electronic Society, WPES 2002, Washington, DC, USA, November 21, 2002*, 2002 1 (cit. on p. 84).
- [130] C. Bartolini and R. Muthuri, "Reconciling Data Protection Rights and Obligations: An Ontology of the Forthcoming EU Regulation", *Proceedings of the Workshop on Language and Semantic Technology for Legal Domain (LST4LD)*, 2015 8 (cit. on p. 90).
- [131] H. J. Pandit, D. Lewis and D. O'Sullivan, *GDPRtEXT - GDPR as a Linked Data Resource*, URL: <https://doi.org/10.5281/zenodo.1146351> (cit. on p. 90).

# Appendix





---

## Low-level Attributes of *Pripolis*

---

In this part, we provide the complete result of low-level attribute classifications using the fine-tuned BERT. In all experiments, the number of epochs are tuned using the validation set. For three attributes (Retention Period, Retention Purpose and Access Type), we also compared the deep learning results with a rule-based classification. The rationale behind this decision was the low number of annotations in the training set. We observed that due to rare number of samples, the ML models are not able to learn the class-specific weights properly. The dataset used in the rule-based experiments, is the same as machine learning one, e.g., 60% of low-level annotations were used to implement the hand-coded rules, 20% of the dataset were used for validating the defined rules and the remaining 20% was kept for the one-time test. Similar to *EULAide*, we used GATE API and JAPE grammar in order to implement linguistic rules. The GATE pipeline contains some common pre-processing steps (tokenisation, sentence splitting, pos tagging) and one transducer with the hand-coded rules. Table A.1 shows some sample rules for finding values of *Retention Period* attribute in *Data Retention* category. We found our rules definitions on experts' annotations. The evaluation conducted by the rule-based approach are encouraging and prove that in the case of scarce training samples, careful definition of phrase heuristic based on the experts' annotations, can produce promising results.

Table A.1: Sample rules for extracting values of Retention Period from Data Retention category.

Rule	Value	Sample
[delete/remove][Token]*[after][number][day/month/year]	Stated Period	1. We remove the entirety of the IP address after 6 months. 2. All stored IP addresses, except the account creation IP address, are deleted after 90 days.
[not][Token]*[delete/remove]	Indefinitely	The posts and content you made will not be automatically deleted as part of the account removal process.
[store/keep/retain/maintain][Token]*[indefinitely]	Indefinitely	1.This data is generally retained indefinitely. 2. The information we collect for statistical analysis and technical improvements is maintained indefinitely.
[store/keep/retain/maintain][Token]*[as long as][Token]+	Limited	1. We will retain your information for as long as your account is active or as needed to provide you services. 2. We will retain your personal information while you have an account and thereafter for as long as we need it for purposes not prohibited by applicable laws
If not one of the above conditions	Unspecified	1. We receive and store certain types of information whenever you interact with us. 2. The personal information collected about you through our online applications and in our communications with you is stored in our internal database.

Value	Rule-Based			BERT-fine-tuned			
	precision	recall	F1	precision	recall	F1	support
Stated Period	1	0.33	0.5	1	0.33	0.5	16
Limited	0.58	0.78	0.67	0.6	0.33	0.43	36
Indefinitely	0.75	0.75	0.75	0	0	0	12
Unspecified	1	0.68	0.81	0.68	0.79	0.73	72
Macro-avg	0.83	0.64	<b>0.68</b>	0.57	0.36	<b>0.41</b>	
Micro-avg	0.8	0.69	<b>0.74</b>	0.68	0.54	<b>0.6</b>	

Table A.2: Retention Period.

Value	Rule-Based			BERT-fine-tuned			
	precision	recall	F1	precision	recall	F1	support
Advertising	0	0	0	0	0	0	3
Analytics/Research	1	1	1	0	0	0	11
Legal requirement	0.8	0.8	0.8	1	0.2	0.33	21
Marketing	1	1	1	0	0	0	4
Perform service	0.53	0.61	0.57	0.79	0.85	0.81	47
Service operation and security	1	0.25	0.4	1	0.25	0.4	17
Unspecified	0.75	0.64	0.69	0.6	0.21	0.32	55
Macro-avg	0.73	0.61	<b>0.64</b>	0.48	0.22	<b>0.27</b>	
Micro-avg	0.7	0.63	<b>0.67</b>	0.76	0.39	<b>0.52</b>	

Table A.3: Retention Purpose.

Value	Rule-Based			BERT-fine-tuned			
	precision	recall	F1	precision	recall	F1	support
Deactivate account	0.67	0.67	0.67	0	0	0	9
Delete account (full)	0.33	0.25	0.29	0	0	0	16
Delete account (partial)	0.86	0.67	0.75	0.5	0.22	0.31	37
Edit information	0.93	0.93	0.93	0.8	0.96	0.87	114
View	0.56	0.69	0.62	0.57	0.31	0.4	51
None	0	0	0	0	0	0	6
Unspecified	0.4	0.5	0.44	0	0	0	19
Macro-avg	0.54	0.53	<b>0.53</b>	0.27	0.21	<b>0.23</b>	
Micro-avg	0.75	0.76	<b>0.76</b>	0.74	0.54	<b>0.62</b>	

Table A.4: Access Type.

Value	P	R	F1	support
Computer information	0.79	0.85	0.82	150
Contact	0.83	0.83	0.83	444
Cookies and tracking elements	0.94	0.96	0.95	341
Demographic	0.82	0.74	0.78	131
Financial	0.89	0.83	0.86	131
Generic personal information	0.7	0.75	0.72	694
Health	1	0.33	0.5	55
IP address and device IDs	0.79	0.88	0.83	173
Location	0.68	0.6	0.64	142
Personal identifier	0.75	0.2	0.32	43
Social media data	0.8	0.57	0.67	35
Survey data	0.5	0.1	0.15	34
User online activities	0.73	0.62	0.67	354
User profile	0.33	0.25	0.29	160
Unspecified	0.73	0.75	0.74	1078
Macro-avg	0.75	0.62	<b>0.65</b>	
Micro-avg	0.76	0.74	<b>0.75</b>	

Table A.5: Personal Information Type.

Value	P	R	F1	support
Collect from user on other websites	1	0.17	0.29	28
Collect in mobile app	0.82	0.69	0.75	71
Collect on mobile website	0	0	0	17
Collect on website	0.67	0.88	0.75	677
Receive from other parts of company/affiliates	0	0	0	22
Receive from other service/third-party (named)	0.62	0.36	0.45	72
Receive from other service/third-party (unnamed)	0.61	0.38	0.47	102
Track user on other websites	1	0.11	0.2	44
Unspecified	0.81	0.55	0.66	742
Macro-avg	0.61	0.35	<b>0.4</b>	
Micro-avg	0.72	0.63	<b>0.67</b>	

Table A.6: Action First Party.

Value	P	R	F1	support
Children	1	0.97	0.99	151
Californians	1	0.92	0.96	58
Citizens from other countries	0.86	0.92	0.89	49
Europeans	1	0.8	0.89	22
Macro-avg	0.96	0.9	<b>0.93</b>	
Micro-avg	0.97	0.94	<b>0.96</b>	

Table A.7: Audience Type.

Value	P	R	F1	support
Collect on first party website/app	0.53	0.51	0.52	134
Receive/Shared with	0.91	0.86	0.89	639
See	1	0.57	0.73	61
Track on first party website/app	0.73	0.63	0.68	118
Unspecified	0.67	0.47	0.55	156
Macro-avg	0.77	0.61	<b>0.67</b>	
Micro-avg	0.82	0.73	<b>0.77</b>	

Table A.8: Action Third Party.

Value	P	R	F1	support
Explicit	0.79	0.79	0.79	387
Implicit	0.72	0.82	0.77	397
Unspecified	0.78	0.7	0.74	543
Macro-avg	0.76	0.77	<b>0.76</b>	
Micro-avg	0.76	0.76	<b>0.76</b>	

Table A.9: Collection Mode.

Value	P	R	F1	support
In case of merger or acquisition	0	0	0	9
Non-privacy relevant change	0	0	0	10
Privacy relevant change	0.87	0.5	0.64	54
Unspecified	0.69	0.96	0.81	109
Macro-avg	0.39	0.36	<b>0.36</b>	
Micro-avg	0.71	0.71	<b>0.71</b>	

Table A.10: Change Type.

Value	P	R	F1	support
Collection	0.48	0.43	0.45	313
First party collection	0.41	0.41	0.41	124
First party use	0.69	0.5	0.58	214
Third party sharing/collection	0.53	0.25	0.34	112
Third party use	0.67	0.11	0.19	59
Both	0	0	0	89
Use	0.45	0.11	0.18	157
Unspecified	0.87	0.89	0.88	1417
Macro-avg	0.51	0.34	<b>0.38</b>	
Micro-avg	0.76	0.66	<b>0.7</b>	

Table A.11: Choice Scope.

Value	P	R	F1	support
Browser/device privacy controls	0.93	0.82	0.87	105
Dont use service/feature	0.52	0.52	0.52	226
First-party privacy controls	0.5	0.27	0.35	75
Opt-in	0.63	0.73	0.67	409
Opt-out link	0.87	0.59	0.7	149
Opt-out via contacting company	0.94	0.68	0.79	119
Third-party privacy controls	0.62	0.48	0.54	86
Unspecified	0.87	0.9	0.88	1284
Macro-avg	0.74	0.62	<b>0.67</b>	
Micro-avg	0.79	0.78	<b>0.78</b>	

Table A.12: Choice Type.

Value	P	R	F1	support
Does	0.98	0.86	0.98	1660
Does Not	0.99	0.84	0.85	223
Macro-avg	0.92	0.92	<b>0.92</b>	
Micro-avg	0.96	0.97	<b>0.97</b>	

Table A.13: Does or Does Not.

Value	P	R	F1	support
Honored	0	0	0	1
Not honored	0.71	1	0.83	23
Mentioned, but unclear if honored	0	0	0	2
Macro-avg	0.24	0.33	<b>0.28</b>	
Micro-avg	0.71	0.83	<b>0.77</b>	

Table A.14: Do Not Track.

Appendix A Low-level Attributes of Pripolis

Value	P	R	F1	support
Generic	0.78	0.86	0.82	156
Data access limitation	0.89	0.5	0.64	61
Privacy review/audit	0	0	0	18
Privacy training	0	0	0	4
Privacy/Security program	1	0.12	0.22	30
Secure data storage	0	0	0	22
Secure data transfer	0.83	0.5	0.62	39
Secure user authentication	0	0	0	17
Unspecified	0	0	0	8
Macro-avg	0.39	0.22	<b>0.26</b>	
Micro-avg	0.81	0.53	<b>0.64</b>	

Table A.15: Security Measure.

Value	P	R	F1	support
Additional service/feature	0.73	0.52	0.61	413
Advertising	0.85	0.83	0.84	301
Analytics/Research	0.86	0.77	0.81	299
Basic service/feature	0.72	0.63	0.67	463
Legal requirement	0.96	0.78	0.86	127
Marketing	0.81	0.61	0.7	367
Merger/Acquisition	1	1	1	59
Personalization/Customization	0.85	0.75	0.8	201
Service operation and security	0.79	0.66	0.71	266
Unspecified	0.86	0.69	0.77	867
Macro-avg	0.84	0.72	<b>0.78</b>	
Micro-avg	0.82	0.68	<b>0.75</b>	

Table A.16: Purpose.

Value	P	R	F1	support
Aggregated or anonymized	0.89	0.92	0.9	220
Identifiable	0.7	0.79	0.74	467
Unspecified	0.95	0.9	0.92	1154
Macro-avg	0.85	0.87	<b>0.86</b>	
Micro-avg	0.88	0.88	<b>0.88</b>	

Table A.17: Identifiability.

Value	P	R	F1	support
User with account	0.79	0.85	0.82	280
User without account	0	0	0	47
Unspecified	0.97	0.98	0.98	1726
Macro-avg	0.59	0.61	<b>0.6</b>	
Micro-avg	0.94	0.95	<b>0.95</b>	

Table A.18: User Type.

Value	P	R	F1	support
General notice in privacy policy	0.77	0.85	0.81	81
General notice on website	1	0.28	0.43	42
No notification	0	0	0	8
Personal notice	0.67	0.4	0.5	38
Unspecified	0.18	0.2	0.19	45
Macro-avg	0.52	0.34	<b>0.39</b>	
Micro-avg	0.62	0.49	<b>0.55</b>	

Table A.19: Notification Type.

Value	P	R	F1	support
None	0.67	0.67	0.67	22
Opt-in	0	0	0	9
Opt-out	0	0	0	5
User participation	0.6	0.5	0.54	22
Unspecified	0.92	0.88	0.9	120
Macro-avg	0.44	0.41	<b>0.42</b>	
Micro-avg	0.83	0.73	<b>0.78</b>	

Table A.20: User Choice.

Value	P	R	F1	support
Profile data	0	0	0	31
Transactional data	0	0	0	15
User account data	0.79	0.79	0.79	100
Other data about user	0	0	0	29
Unspecified	0.67	0.45	0.54	85
Macro-avg	0.29	0.25	<b>0.27</b>	
Micro-avg	0.74	0.45	<b>0.56</b>	

Table A.21: Access Scope.

Value	P	R	F1	support
Named third party	0.8	0.68	0.74	415
Other part of company/affiliate	0.8	0.4	0.53	116
Other users	1	0.57	0.73	29
Public	1	0.67	0.8	44
Unnamed third party	0.77	0.89	0.83	596
Unspecified	0.74	0.53	0.62	182
Macro-avg	0.85	0.63	<b>0.71</b>	
Micro-avg	0.79	0.72	<b>0.75</b>	

Table A.22: Third Party Entity.

# List of Figures

---

1.1	Excerpts from <i>ResearchGate</i> contractual agreements. . . . .	2
1.2	The high-level workflow of our approach. . . . .	7
2.1	An example of RDF graph. Resources are denoted by circles and literals are denoted by rectangles. . . . .	14
2.2	A fully connected feedforward multilayers neural network. . . . .	21
2.3	Cosine similarity. . . . .	23
2.4	A Convolutional Neural Network (CNN) architecture for sentence classification [9]. . . . .	23
3.1	Tldrlegal service showing MIT license summary. . . . .	26
3.2	Data Rights Finder service showing analysis of PayPal privacy policy. . . . .	27
3.3	GRCTC's developed tool for querying the semantically enriched text. . . . .	29
3.4	Evaluation of ResearchGate privacy policy using the <i>Privacycheck</i> Chrome extension. . . . .	34
3.5	PrivacyGuide snapshots. . . . .	35
4.1	Architecture of the <i>EULAide</i> system. . . . .	41
4.2	GATE EULA OBIE Pipeline. . . . .	42
4.3	<i>EULAide</i> platform Web interface showing the permission, duty and prohibition clusters for a user provided EULA. . . . .	47
4.4	A sample mockup of a mobile app communicating to the <i>EULAide</i> Web service. . . . .	48
5.1	<b>The OPP-115 dataset.</b> The top level of the hierarchy (shaded blocks) defines high-level categories. The lower level defines a set of privacy attributes, each assuming a set of values. . . . .	61
5.2	An example of annotations by an expert. . . . .	61
5.3	CNN architecture for multi-label classification of privacy policies. . . . .	64
5.4	<i>Pripolis</i> Web interface showing the predicted (high/low)-level classes for a user provided privacy policy. . . . .	66
5.5	Disagreement example for the Amazon privacy notice. . . . .	71
6.1	GDPR attention in 2018, published by the EC. . . . .	76
6.2	Architecture and workflow of <i>KnIGHT</i> . . . . .	78



# List of Tables

---

4.1	Vocabularies and ontologies for EULAs. . . . .	41
4.2	Example of a <code>Permission</code> as extracted by <i>EULAide</i> . . . . .	44
4.3	Example of annotated <code>permissions</code> in Apache. . . . .	45
4.4	Example of features extraction. . . . .	45
4.5	Specification of end-user license agreements. . . . .	49
4.6	Lenient IAA for two annotators. . . . .	50
4.7	Evaluation of OBIE pipeline. . . . .	51
4.8	Total extracted instances & machine-generated clusters with (Mf) and without (M) considering features. . . . .	53
4.9	Rand index for 5 humans (h1-h5) and machine-generated clusters with (Mf) & without (M) features. . . . .	54
4.10	Average results. . . . .	54
4.11	Average time (in seconds). . . . .	56
4.12	Average percentage of questions results (%). . . . .	56
4.13	Average scores of six participants for the usability questionnaire (Max=7). . . . .	57
5.1	<i>Disconnect</i> privacy icons with their descriptions & <i>Polisis</i> 's interpretation from Harkous et. al. [67]. . . . .	67
5.2	Label distribution in the two gold standards for the high-level categories; Tr:Train; V:Validation; T:Test. . . . .	68
5.3	F1 for three models on the two gold standards in (%) with tuned epochs on validation; Threshold=0.5; V:Validation; T:Test. . . . .	69
5.4	Macro averages on the union-based gold standard in (%) with tuned epochs on validation; Threshold=0.5; V:Validation; T:Test. . . . .	70
5.5	Accuracy of risk color prediction for the five privacy icons; R:Red; G:Green; Y:Yellow. . . . .	72
6.1	Potential mappings between a privacy policy and the GDPR, based on our observation. . . . .	77
6.2	Posteriori assessment by 4 experts (E1-E4) for four privacy policies. . . . .	82
6.3	Pair-wise agreement between experts. . . . .	82
6.4	Pair-wise weighted kappa between experts. . . . .	82
6.5	Example of detected links & experts (E1-E4) Assessments (R: related, P: partially related, U:unrelated). . . . .	83
6.6	Average F-measure & total time of 2 regular end-users annotations for 4 privacy policies. . . . .	84
A.1	Sample rules for extracting values of Retention Period from Data Retention category. . . . .	106
A.2	Retention Period. . . . .	106
A.3	Retention Purpose. . . . .	106
A.4	Access Type. . . . .	106
A.5	Personal Information Type. . . . .	107

*List of Tables*

---

A.6 Action First Party. . . . .	107
A.7 Audience Type. . . . .	107
A.8 Action Third Party. . . . .	107
A.9 Collection Mode. . . . .	107
A.10 Change Type. . . . .	107
A.11 Choice Scope. . . . .	107
A.12 Choice Type. . . . .	107
A.13 Does or Does Not. . . . .	107
A.14 Do Not Track. . . . .	107
A.15 Security Measure. . . . .	108
A.16 Purpose. . . . .	108
A.17 Identifiability. . . . .	108
A.18 User Type. . . . .	108
A.19 Notification Type. . . . .	108
A.20 User Choice. . . . .	108
A.21 Access Scope. . . . .	108
A.22 Third Party Entity. . . . .	108