CUSTOMER-CENTERED LEGALTECH: AUTOMATED ANALYSIS OF STANDARD FORM CONTRACTS

Daniel Braun / Elena Scepankova / Patrick Holl / Florian Matthes

Research Associates, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
daniel.braun@tum.de / elena.scepankova@tum.de / patrick.holl@tum.de; https://wwwmatthes.in.tum.de

Professor, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
matthes@tum.de; https://wwwmatthes.in.tum.de

Abstract: *Customers are regularly confronted with complex standard form contracts, especially online. Most of the time, customers do not read these contracts at all and even if they try, they are not always able to understand them. In this paper, we present a technological approach to address this imbalance of powers between customers and companies and a prototypically implementation which analyses, assesses, and summarizes terms of services from German web shops.*

1. Introduction

So-called «LegalTech» solutions aim to use technologies, e.g. from the field of artificial intelligence, to support or automate processes within the legal domain. Such solutions could help to make legal expertise more accessible by lowering current cost and time expenditure. However, so far, many LegalTech solutions focus on corporate users due to economic and practical considerations: the prevalent imbalance of powers between companies and consumers, which exists for centuries (cf. Section 4).

Standard form contracts trace back to the 19th century. In the age of industrialisation, entering into contracts has been accompanied by the unilateral use of pre-formulated rules tailored to one party's own interests and thus resulting in an imbalance of powers between the contracting parties [ZERRES, 2014]. Today, customers are confronted with standard form contracts every day in form of Terms of Services (ToS), for example when they buy something online or register for an online service. Studies with more than 45'000 participants have shown that only 0.1% to 0.2% of customers read ToS of online shops [BAKOS ET AL., 2014]. While standard form contracts regularly reflect an imbalance of contracting power, this imbalance is even stronger in situations where one of the contracting parties is a consumer without a professional legal background and the other one is a company with a potentially huge legal department. The relevance of this is visible by the amount of jurisprudence in this area, comprising more than 28'000 judgments in Germany only [JURIS, 2017]. Therefore, in this paper, we present an approach to use Artificial Intelligence (AI) to empower customers of web shops to make educated decisions about where to buy or not by automatically identifying Terms of Services of German web shops, assessing their content regarding lawfulness and customer friendliness, as well as summarising them in a simplified language in order to make them understandable for non-legal-experts.

2. Related Work

A similar goal is pursued by the project «Terms of Service; Didn't Read» (ToS;DR) [BINNS AND MATTHEWS, 2014]. Instead of automatically assessing and evaluating the ToS, ToS;DR is crowd-sourced and provides

manually generated summarisations of ToS from many major websites. However, the fact that ToS;DR is crowd-sourced affects their scalability and topicality.

An automated approach to analyse online standard form contracts was presented by LIPPI ET AL. Their analysis focuses solely on so-called «unfair clauses» which are forbidden under the law of the European Union [LIPPI ET AL., 2017]. In their experiment, they analysed Terms of Services from 20 major websites regarding eight unfair clauses. In a leave-one-document-out evaluation, they achieved a precision of 0.62 using a Support Vector Machine. In contrast to our approach, LIPPI ET AL. try to do a binary classification (unfair clause exists / does not exist), while we try to gather additional information and summarise them.

In order to create these summaries, a system first has to obtain the relevant information from the text. Information Retrieval (IR) for legal texts has gained a lot of attraction in recent years. Examples are MCCALLUM [2015], GRABMAIR ET AL. [2015], and FRANCESCONI ET AL. [2010], or, for German texts, WALTER AND PINKAL [2006], and WALTL ET AL. [2010]. The issue of simplifying legal texts was, e.g. addressed by BHATIA [1983] and COLLANTES ET AL. [2015].

3. Customer Perspective

As mentioned before, only 0.1% to 0.2% of customers read the ToS of online shops [BAKOS ET AL., 2014]. In acknowledgement of this fact, the European lawmaker has limited the creative leeway for companies, when it comes to standard form consumer contracts.[1] One might ask oneself why consumers should care about unlawful clauses in ToS, because in case of a dispute, these clauses are void anyway. In reality however, at least for online shopping, the amount in dispute is often so low that consumers avoid legal steps, even if they are in the right. If customers are aware of unlawful clauses before making a purchase, they would most possibly consider different vendors.

Moreover, customers tend to have individual demands, e.g. regarding privacy, which surpass legal regulations. A tool which automatically analyses ToS not only regarding their lawfulness, but also regarding individual user-defined criteria, could help customers in making more educated decisions about where to buy (or register) or not. Such a tool does not replace or automatize legal advice allowed to be given exclusively by legal professionals. It, however, serves as a clarifying tool with respect to legal terms and hence directly fosters the constitutional principle of Legal Clarity.[2]

Interestingly enough, this also applies to one of the few existing LegalTech solutions which focuses on consumers, the service Flightright[3]. While LegalTech for companies seems to be focused on automatising or at least supporting existing legal processes, customer-centered LegalTech seems to be focused on providing advice and support in situations where it was previously neither available, nor economically accessible.

4. Legal Perspective

4.1. Imbalance of powers

When we consider Standard Form Contracts from a consumer perspective, we usually encounter an imbalance of powers between the parties. In contrast to the company, the consumer is exposed to specific risks as a market participant. It is assumed, that there are divergent economic interests between consumers and companies [KOCHER, 2007] which result in a structural imbalance of powers.[4] Divergent interests themselves, however, do not constitute this imbalance of powers; it is the further existence of a negotiation disparity between those two groups of market participants, which leads to this situation. As the free market does not provide the

---

[1] Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts, OJ L 95/29 of 21 April 1993.
[2] Art. 20 Abs. 3 GG (Grundgesetz – German Constitution), https://www.gesetze-im-internet.de/englisch_gg/index.html (all websites last visited in January 2018).
[3] https://www.flightright.de/.
[4] Cf. German Constitutional Court BVerfGE 89, 214 of 19 October 1993.

necessary self-regulation, the state needs to intervene by providing a set of legal regulations to address the mentioned imbalance [SCHIRMBACHER, 2005].

4.2. Reasons

In German law, it is well-established opinion that the consumer has an inferior market position in comparison to the company [ZIMMERMANN, 2005]. The reasons for this situation are manifold. A first reason is seen in the sociologically caused asymmetries between the group of consumers and the one of the companies. These asymmetries are based on intellectual, psychologic and economic imbalances between the parties. Secondly, it is acknowledged that the group of the consumers is more inhomogeneous than the one of the companies [BEATER, 2000]. This results in a comparably diffuse state of interests among the consumers, which leads to a relatively weak «group consciousness» in the group of consumers [KOCHER, 2007]. Finally – which in our opinion is the most relevant aspect – the reason is seen in the «typeable information deficit» of the consumers compared to companies [FLEISCHER, 2001]. This means that companies typically have more information about the market, the products and the behaviour thereof than consumers. This information deficit fundamentally impacts on the imbalance of powers. With SaToS we try to put the consumer into a more informed position and thus directly address one of the reasons for the existant imbalance.

4.3. «SaToS» approach

The consumer protection law aims to protect the consumer and allow for «optimal market decisions» [OEHLER, 2006]. This aim to enable consumers to take «optimal market decisions» is considered as the fundamental and is intended to foster the faith of the consumers into market processes [TAMM, 2011]. German consumer law uses different regulatory instruments to achieve this, by preventively designing consumer protection law and by subsequently declaring contractual clauses not in accordance with those legal provisions as «void». In order to work effectively, legal regulations comprise both market-complementary as well as market-compensatory instruments [REICHARDT, 2008].

With SaToS we likewise aim to enable the consumer to take «optimal market decisions». We consider both above mentioned ways as important, but we think that there are obstacles which hinder legal regulations to achieve full value. The most important fact is that reading and understanding legal contracts is hindered by the fact that in comparison to regular language, legal language is characterised by a high degree of abstractness. In order to fulfil its function as a merit instance for any socially relevant behaviour, law itself needs to guard its capability of abstractly reacting even to unforeseen situations, which results in formulations characterised by a low level of comprehensibility. By summarizing contractual terms and translating their legal and linguistic complexity into a simplified language (summary generation), with SaToS' first functionality we provide the customer with the possibility to understand his rights and duties and to take decisions based on knowledge, not on – justified or unjustified – trust towards the shop provider.

Secondly, standard form contracts like Terms of Services address fundamental conditions of performance of a certain business. In the context of online shopping, they set the provisions for e.g. payment, delivery, revocation or liability. Due to the abstract character of the legal language stipulating contractual rights and duties on the one hand, and normative requirements in legal regulations and judicial decisions on the other hand, the consumer often finds himself not capable to understand and assess the validity or invalidity of his/her contract. With our second functionality, we thus automatically (a) identify the differences between the clauses of (online shops as) companies and (b) assess their (un)lawfulness.

4.4. Act of Out-of-Court Legal Services

Concerning our functionality to automatically identify unlawful clauses, we are aware of the German Act of Out-of-Court Legal Services[5] which stipulates that in individual cases legal advice against payment shall only be provided by legal personnel, e.g. lawyers, legal counsels, tax consultants. Our functionality, however, serves as a mere clarification tool operated on standard form contracts in a general way. As we neither intend nor are capable of providing specific legal advice in individual cases, our tool focuses on enhancing the understanding of legal language. Moreover, we intend to cooperate with national Consumer Protection Agencies, which have already expressed interest in using our application. Those institutions are entitled to give out-of-court legal advice within their area of expertise and responsibility.[6]

5. Technical Approach

As a proof of concept, we developed a research prototype called SaToS (software aided analysis of Terms of Services) which is able to automatically identify ToS pages of German web shops, summarise them in a simplified language and assess them with regard to their lawfulness and customer friendliness. So far, the prototype is limited to clauses addressing warranty and the right of withdrawal. A screenshot of the prototype is shown in Figure 1.
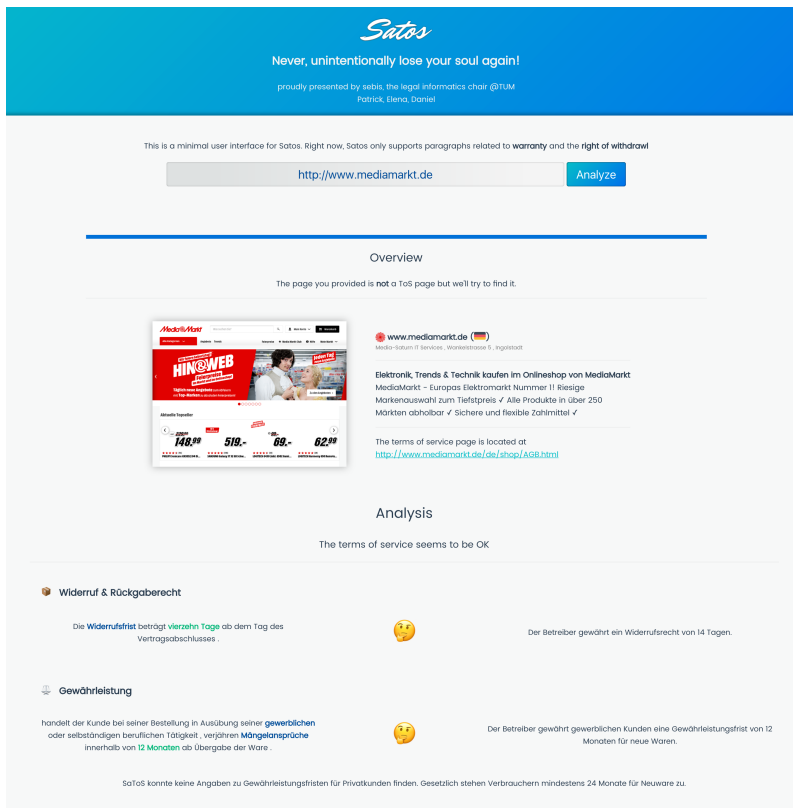

Figure 1: Screenshot research prototype

---

5   https://www.gesetze-im-internet.de/englisch_rdg/index.html.
6   Cf. § 8 Abs. 1 Nr. 4 Act of Out-of-Court Legal Services.

Before we can start with the actual analysis and summarisation, we first have to find the Terms of Services of a web shop. In order to be a user-friendly tool, the tool should be able to automatically find the Terms of Services page and not require the user to enter the URL manually. SaToS uses a hybrid approach, combining a rule system and machine learning, to find ToS pages reliably and fast. In a first step, we analyse all URL according to a set of rules. One common pattern we identified is that the ToS URL often contain «AGB». The classifier separates the URL strings into the following components: scheme specifier, network location part, path, query parameters. The path and query parameters are matched against a set of pre-defined, weighted rules. If the matches reach a certain threshold, we consider that a given URL points to a potential ToS page. With this approach we can classify a URL within less than 0.001s and with a precision of 0.7953 (cf. Section 6 for a more detailed evaluation). For those cases in which we cannot find a ToS URL using the rule-based approach, we use a Naive Bayes classifier as fallback solution. While the classification with the machine learning approach is significantly slower, it also has a much higher accuracy of 0.9115.

Once the ToS have been found, from a computer science perspective, there are three main challenges which have to be solved: extraction of the information, assessment of the lawfulness / customer friendliness and summary generation. Figure 2 shows the pipes and filters architecture which we use to address all of these tasks.
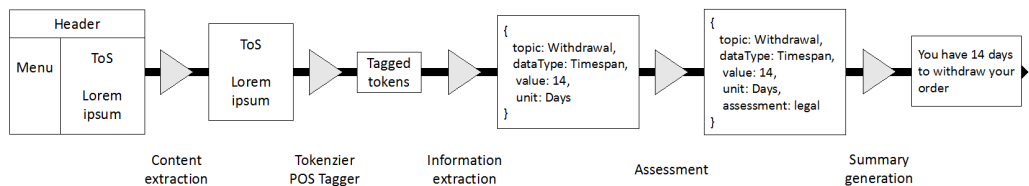
Figure 2: Pipes and filters architecture of the prototype

In order to extract the necessary information from the Terms of Services, we first have to extract the content of the ToS from the rest of the website, i.e. remove the navigation, header, footer and so on. Afterwards, we use standard natural language processing technologies (sentence splitter, tokenizer, stemmer, POS tagger) to annotate the text with linguistic information.

This stripped and annotated version of the text can then be used to extract relevant information (like the time of withdrawal). Currently, our system uses a rule-based approach for extraction. Some of the rules we developed to extract the necessary information are shown in Table 1.

Topic	Rules (translated from German)
Right of warranty	– warranty . . . ([0-9]* | [one | two | . . .]) [day(s) | month(s) | year(s)] AND used OR NOT used (goods | products) -warranty . . . used (goods | products) . . . excluded
Right of withdrawal	– product . . . original (packaging | packed | . . .) . . . (return | send back)– original (packaging | packed | . . .) . . . (return | send back)
Period of withdrawal	– withdraw . . . ([0-9]* | [one | two | . . .]) [day(s) | month(s) | year(s)]
Obligation to inspect product	– warranty . . . [inspect|report] AND NOT merchant
Risk of loss	– [risk of loss | bearing the risk] . . . [shipped | carriage of goods] . . . consumer

Table 1: Extraction rules for ToS (excerpt)

After the extraction is finished, the information is stored in a structured format, as shown in Figure 2. In order to assess this information, we need a standard to compare them to. For this purpose, we built a knowledge base containing a structured representation of important laws and judgements regarding ToS of web shops. We use the data from the knowledge base as reference and generate the assessment accordingly. An assessment can either be that the text is unlawful, lawful, or customer friendly, if it surpasses the legal standard in a customer-friendly manner, e.g. by providing a period of withdrawal longer than 14 days.

Finally, the now structured information is translated back to a simplified text version, using a template language and SimpleNLG [GATT AND REITER, 2009] as surface realiser.

In order to make the reasoning of our tool transparent and comprehensible, we do not only provide the assessment and summarisation to the user, but also the original text part, on which the assessment and summarisation are based, and highlight which pieces of information were taken into account (cf. Figure 1, left side).

Moreover, in case we do not find any specific information about a topic in the Terms of Services, we inform the user about his general rights according to current legal standards, e.g. a warranty period of at least 24 months for new goods for consumers.

6. Evaluation

As mentioned before, we use a hybrid approach of client-based rules and server-based ML. Since all further analyses are based on the correct classification of ToS pages, we conducted an evaluation of both components.

6.1. ToS Classification

We collected a dataset of 3'424 URLs. 2'592 from ToS pages, manually labeled by a price comparison website, and 832 from other webshop pages. We split the dataset into training (200 ToS and 200 Other) and test (2'392 ToS and 632 Other). The results of the evaluation are shown in Table 2. It is obvious, that the ML approach performed significantly better, with regard to precision, recall, and F-score. Given the fact, that the ML classifier was trained with a relatively small, non-optimized dataset, the results are promising, keeping in mind that we use an online learning approach and expect the system to improve over time. One might wonder, why one should use a hybrid approach, although ML performs better in every category. The fifth column in Table 2 shows the average time in seconds, that was needed to classify a URL. If successful, the rule-based approach is significantly faster and hence preferable.

Approach	Precision	Recall	F-score	Øt in s
ML	0.9115	0.8219	0.8644	1.435
Rule-based	0.7953	0.5393	0.6428	0.001

Table 2: Evaluation ToS Classification

6.2. Information Extraction

In order to get a first idea about the quality of the information extraction of our prototype, we randomly picked 20 ToS from the above mentioned dataset of ToS pages. We asked a legal professional to manually extract the information, which our current prototype is able to extract (i.e. right of withdrawal and warranty period). In this way, for each ToS up to five pieces of information were manually extracted (time of withdrawal, warranty period for consumer new good, warranty period of consumer used good, warranty period of tradesperson new good, warranty period of tradesperson used good). Afterwards, we ran our tool on the same ToS and compared the results. The result of the comparison is shown in Table 3. If the manually and automatically extracted information is equal, the information was correctly extracted. If the automatically extracted information differs from the manually extracted information (e.g. withdrawal period of 14 instead of 30 days), it is classified as wrongly extracted information. If information was extracted manually for a topic but not automatically, it is classified as missing information. In no case, information that was not extracted manually was extracted automatically.

Topic	Correctly extracted information	Wrongly extracted information	Missing information
Withdrawal	18	1	0
Warranty	41	3	4

Table 3: Evaluation Information Extraction

7. Conclusion

The evaluation of our research prototype SaTos, which automatically detects, summarises, and analyses ToS from German webshops regarding their lawfulness and customer friendliness, has shown, that LegalTech can indeed help to empower customers and address the imbalance of powers between consumers and large companies. While the prototype is still at an early stage, the first evaluation looks promising and it gives a glance at the possibilities LegalTech can provide to customers.

Both our functionalities increase the information level of the consumers and thus fundamentally influence the imbalance of powers. With SaToS we are able to address one of the major reasons for the existence of consumer protection – the information deficit – directly and more specifically than with the existant broad regulation system.

Eventually, one may also expect that adding transparency into the legal regulation jungle will have an impact on companies' approach towards this topic. It may turn advisable for companies to consider a clearer regulation drafting right from the start instead of being highlighted expressively as «unlawful» or «customer-unfriendly» by an independent third party. This is why our solution might contribute to a shift in legal regulation formulation in general.

8. References

BAKOS, YANNIS/MAROTTA-WURGLER, FLORENCIA/ TROSSEN, DAVID, Does anyone read the fine print? Consumer attention to standard-form contracts, The Journal of Legal Studies, 2014, Issue 43, p. 1–35.

BEATER, AXEL, Verbraucherschutz und Schutzzweckdenken im Wettbewerbsrecht, Mohr Siebeck, 2000, p. 117.

BHATIA, VIJAY, Simplification v. easification-the case of legal texts. Applied linguistics, Volume 4:42, 1983.

BINNS, REUBEN/MATTHEWS, DAVID, Community structure for efficient information flow in «tos; dr», a social machine for parsing legalese, In Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, p. 881–884.

COLLANTES, MIGUEL/HIPE, MAUREEN/SORILLA, JUAN LORENZO/TOLENTINO, LAURENZ/SAMSON, BRIANE, SIMPATICO, A text simplification system for senate and house bills, In Proceedings of the 11th National Natural Language Processing Research Symposium, 2015, p. 26–32.

FLEISCHER, HOLGER, Informationsasymmetrie im Vertragsrecht, Doctoral dissertation, Universität zu Köln, 2001.

FRANCESCONI, ENRICO/ MONTEMAGNI, SIMONETTA/PETERS, WIM/TISCORNIA, DANIELA, Semantic processing of legal texts: Where the language of law meets the law of language, Volume 6036. Springer, 2010.

GATT, ALBERT/REITER, EHUD, SimpleNLG: A realisation engine for practical applications, Proceedings of the 12th European Workshop on Natural Language Generation, Association for Computational Lingustics, 2009.

GRABMAIR, MATTHIAS/ASHLEY, KEVIN/CHEN, RAN/SURESHKUMAR, PREETHI/WANG, CHEN/NYBERG, ERIC/WALKER, VERN, Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools, In Proceedings of the 15th International Conference on Artificial Intelligence and Law, ACM, 2015, p. 69–78.

JURIS RECHTSINFORMATIONSDATENBANK, https://www.juris.de/r3/search, last visited on 29 October 2017.

KOCHER, EVA, Funktionen der Rechtsprechung, Mohr Siebeck, 2007, p. 48.

LIPPI, MARCO/PALKA, PRZEMYSLAW/CONTISSA, GIUSEPPE/LAGIOIA, FRANCESCA/MICKLITZ, HANS-WOLFGANG/PANAGIS, YANNIS/SARTOR, GIOVANNI/TORRONI, PAOLO, Automated Detection of Unfair Clauses in Online Consumer Contracts, Legal Knowledge and Information Systems, 2017, p. 145–154.

MCCALLUM, ANDREW, Information extraction: Distilling structured data from unstructured text, Queue, Issue 3(9), 2005, p. 48–57.

OEHLER, ANDREAS, Verbraucher und Recht (VuR), Nomos, 2006, p. 294.

REICHARDT, VANESSA, Der Verbraucher und seine variable Rolle im Wirtschaftsverkehr, Duncker & Humblot, 2008, p. 47.

SCHIRMBACHER, MARTIN, Verbrauchervertriebsrecht, Nomos, 2005, p. 124.

TAMM, MARINA, Verbraucherschutzrecht, Mohr Siebeck, 2011, p. 19.

WALTER, STEPHAN/PINKAL, MANFRED, Automatic extraction of definitions from german court decisions, In Proceedings of the workshop on information extraction beyond the document, ACL, 2006, p. 20–28.

ZERRES, THOMAS, Principles of the German law on standard terms of contract, Jurawelt, 2014.

WALTL, BERNHARD/LANDTHALER, JÖRG/SCEPANKOVA, ELENA/MATTHES, FLORIAN/GEIGER, THOMAS/STOCKER, CHRISTOPH/SCHNEIDER, CHRISTIAN, Automated Extraction of Semantic Information from German Legal Documents. In: Schweighofer, Erich/Kummer, Franz/Hötzendorfer, Walter/Sorge, Christoph (Eds.), Trends and Communities of Legal Informatics, Tagungsband des 20. Internationalen Rechtsinformatik Symposions IRIS 2017, books@OCG, Wien 2017, p. 217–224.

ZIMMERMANN, REINHARD, The New German Law of Obligations, Oxford University Press, 2005, p. 165.