NLP for Consumer Protection: Battling Illegal Clauses in German Terms and Conditions in Online Shopping

Daniel Braun
Technical University of Munich
Department of Informatics
`daniel.braun@tum.de`

Florian Matthes
Technical University of Munich
Department of Informatics
`matthes@tum.de`

Abstract

Online shopping is an ever more important part of the global consumer economy, not just in times of a pandemic. When we place an order online as consumers, we regularly agree to the so-called "Terms and Conditions" (T&C), a contract unilaterally drafted by the seller. Often, consumers do not read these contracts and unwittingly agree to unfavourable and often void terms. Government and non-government organisations (NGOs) for consumer protection battle such terms on behalf of consumers, who often hesitate to take on legal actions themselves. However, the growing number of online shops and a lack of funding makes it increasingly difficult for such organisations to monitor the market effectively. This paper describes how Natural Language Processing (NLP) can be applied to support consumer advocates in their efforts to protect consumers. Together with two NGOs from Germany, we developed an NLP-based application that legally assesses clauses in T&C from German online shops under the European Union's (EU) jurisdiction. We report that we could achieve an accuracy of 0.9 in the detection of void clauses by fine-tuning a pre-trained German BERT model. The approach is currently used by two NGOs and has already helped to challenge void clauses in T&C.

1 Introduction

NLP, and technology more broadly, has improved the access to knowledge in many domains. It is no longer necessary to pay thousands of dollars for a lexicon like the Encyclopædia Britannica or to hire a translator to understand texts in other languages. The legal domain is arguably one of the biggest resistance to digitisation efforts. While, in some aspects, it still struggles to catch up with other industries, technology has started to change the landscape of legal service provision. So far, consumers rarely benefit from this development. On the contrary, mostly big companies and law firms benefit. Most of the existing so-called "LegalTech" tools, like Lexis Advance[1], Lexical Labs[2], and ANVI [3], to name just a few, are tailored to the needs of companies and law firms, rather than consumers and consumer protection agencies. Thereby, Legal-Tech tools are not only missing the opportunity to democratise access to legal advice, by making it more affordable and available, they are actively increasing the current imbalance of power between companies and consumers.

In this paper, we describe, how we apply NLP technology to automatically assess clauses in German T&C from consumer online shops, to find void clauses and help to protect consumers from them. Unlike the, relatively little, existing work (see Section 2), we focus on organisations that represent consumer interests as users. By focusing on such organisations, rather than individual consumers, we hope to be able to increase the impact of our work. While tools for individual consumers usually only benefit those who are using them, consumer protection agencies legally challenge void T&C they find, forcing their change and hence benefiting all consumers. We also believe that the task of ensuring that companies adhere to consumer contract and distance selling laws should not be left to consumers alone.

2 Related Work

As mentioned before, the existing research in the area of the legal analysis of T&C focuses on individual consumers as users.

The project "Terms of Service; Didn't Read" (ToS;DR) from Binns and Matthews (2014) uses

---

crowd-sourcing to provide manually generated summarisations of the ToS from many major online platforms, like Facebook and Twitter. However, the fact that ToS;DR is crowd-sourced affects the scalability and topicality of the project.

The SaToS project (Software-aided analysis of Terms of Services) (Braun et al., 2017, 2018, 2019a,b) automatically summarise and assess T&C for consumers using dependency parsing and other rule-based approaches, however, only covering a few selected aspects of T&C.

CLAUDETTE is a project at the European University Institute (Micklitz et al., 2017; Lippi et al., 2017; Contissa et al., 2018b,a; Lippi et al., 2019b,a,c; Liepina et al., 2019) which focuses on the detection of unfair clauses in terms of the legislation of the EU. Originally focused on Terms of Services from tech giants like Netflix, Google, Microsoft, and Snapchat, CLAUDETTE now mainly focuses on the analysis of privacy policies.

Since the introduction of the General Data Protection Regulation (GDPR) in the EU, the interest in the analysis of privacy policies has increased in general (see e.g. Harkous et al. (2018) and Torre et al. (2020)).

3 The Role of NGOs in Consumer Protection

The folk wisdom that being right does not automatically lead to getting justice is specifically true for the area of consumer protection, where there is regularly a strong imbalance of power between the involved parties, a single consumer on one side and a potentially large corporation on the other side. In acknowledgement of this fact, many legislators have given NGOs in the area of consumer protection special and extensive rights to assist and represent consumers and their interests. At the same time, consumer advocates and consumer protection agencies are chronically underfunded in many countries. With their limited financial means, consumer advocates all over Europe struggle to keep up with the demand generated by the increasing importance of digital offerings. In 2018, the consumer protection agencies in Germany received in total 184,579 complaints from consumers. 65,370 of these complaints (more than 35%) were related to digital offerings. In comparison, only 36,945 complaints (20%) were received about products and services from the financial industry (Verbraucherzentrale Bundesverband e.V., 2019). In addition to providing individual counselling to consumers, consumer advocates increasingly try to monitor (digital) markets proactively and react to negative developments before consumers are harmed. Monitoring markets as big as eCommerce and proactively act against void clauses in standard form contracts is, at scale, simply not possible without automation of the underlying processes.

For the work presented in this paper, we collaborated with two consumer protection NGOs from two different German states, which are mainly funded by the government and enjoy special privileges when it comes to taking legal actions on behalf of consumers. We worked with five legal experts from these organisations over a period of three years, from 2017 to 2020.

4 Data Corpus

Building a corpus for the automated legal assessment of T&C is far from trivial. On the one hand, we want to have a realistic distribution of clauses in our corpus, with regard to their legality and topics, on the other hand, we need a sufficient number of void clauses in order to be able to train statistical classification models. If we would only use complete T&C, we would need thousands of contracts to find a sufficient number of void clauses.

4.1 Sources

We, therefore, decided to combine three approaches for gathering data:

- We took 78 clauses from a database that is maintained by the organisations we collaborated with. This database contains clauses that have been successfully challenged legally by the organisations and are therefore void.

- We randomly selected 24 complete T&C from the corpus provided by Braun and Matthes (2020), which together consist of 968 clauses.

- The experts actively searched on the internet for clauses about topics they identified as specifically relevant for their everyday work and also specifically for void clauses from these topics. Additional 140 clauses were collected in this way.

Overall, the corpus consists of 1,186 clauses. On average, a clause in our corpus consists of more than 55 words.

Since contracts, under German law, are protected by copyright, we are not allowed to publish the corpus. However, it can be shared on request for non-commercial, scientific purposes.

4.2 Annotation

The 78 clauses which were extracted from the existing database were not manually labelled, because they already have been classified as void by successful legal proceedings.

For all other clauses in the corpus, we had each clause labelled independently by two experts with (potentially) "void" or "valid". Generally speaking, a contract clause is void, if it contains a regulation that violates the law. The final decision of whether a clause is void or not, can, therefore, only be made by a court of law. However, given their expertise and experience in consumer protection law, the experts we worked with can make reasonable assumptions about whether or not a given clause could be ruled to be void, based on the law and existing court decisions.

Some German laws governing the drafting of T&C contain very specific regulations. For example, §355 No. 2 of the German civil code (Bürgerliches Gesetzbuch, BGB) states that *"The withdrawal period is 14 days."* All clauses providing less than 14 days of withdrawal period for consumers are therefore void. Other regulations, however, are more vague. §307 No. 1 BGB, for example, states that clauses are void, if *"[...] they unreasonably disadvantage the other party to the contract [...]"*. Such vague terms need to be interpreted, e.g. by court decisions or legal literature. Therefore, we asked the experts to shortly justify each of their assessment in a commentary and give references to laws or court decisions where appropriate. We then compared the annotations and provided the experts with a list of the conflicting annotations, which they then resolved together by agreeing on one common assessment.

We found the old prejudice of "two lawyers, three opinions" to carry a certain amount of truth. The inter-annotator agreement (before the resolution phase) was between 76% (for the annotation of complete T&C) and 64% (for the annotation of the hand-picked clauses).

4.3 Analysis

Table 1 shows which topics the clauses in the corpus cover and how many clauses for each topic are void. Since a clause can belong to multiple topics, the sum of the counts is larger than the number of clauses. The numbers are also not representative, since the experts actively searched for (void) clauses covering specific topics. The fact that more than 41% of all payment clauses were void, but just about 12% of all delivery clauses, hence, gives no indication about whether payment clauses are generally more likely to be void.

Therefore, we want to focus only on data from T&C that were annotated completely for a moment, because they provide a more realistic picture of the situation. The experts annotated 24 complete T&C. In these 24 T&C, they found 73 void clauses, about three clauses per contract. The contracts consist of 50 clauses per contract on average, which means that about 6% of all clauses are void. The experts were surprised that the ratio of void clauses is that high. They said they never before analysed all aspects of such a large number of T&C and would not have expected to find so many void clauses, and also decided to take actions about some of the clauses they found during the annotation process. So already at this stage, our work had a (small) impact and helped to protect consumers better.

Many void clauses differ only in relatively small aspects from their valid counterparts. A clause about default interest, for example, becomes void if the default interest is set at six percentage points above the base interest rate, instead of five percentage points. The clause *"In the event of a default in payment by the buyer, the seller is entitled to charge interest on the amount outstanding at the rate of six percentage points above the central bank rate at the time payment is due."*, would therefore be void. Such clauses are, linguistically, almost identical. However, there are also a few types of clauses, e.g. defining automatic price increases for subscriptions, that are virtually always void in the data set, independent from the individual phrasing of the clause.

It should be noted that the data in Table 1 only covers clauses that were present and void. In cases of an existing information obligation, the absence of a specific clause might also be unlawful. The fact that the corpus includes 24 T&C, but we found only 18 arbitration clauses imply that at least six companies may not have fulfilled their legal obligation to inform consumers about the EU Online Dispute Resolution (ODR) platform (European Parliament and Council of the European Union, 2013).

Topic	#clauses	#void
minimum age	12	0
applicability	22	1
applicable law	12	1
arbitration	18	1
changes	3	0
conclusion of contract	135	8
delivery	117	14
description	8	0
disposal	8	0
intellectual property	21	0
language	9	1
liability	99	43
party	26	0
payment	305	126
personal data	64	1
place of jurisdiction	11	2
prices	38	9
retention of title	26	4
severability	13	6
text storage	10	0
warranty	43	9
withdrawal	209	26

Table 1: Distribution of clause topics and void clauses in the corpus

5 Approach

The BERT language model (Devlin et al., 2019) has been shown to be effective on a wide range of tasks in the legal domain, including Named Entity Recognition (Chalkidis et al., 2020), annotation of legal concepts (Chalkidis et al., 2020), and evidence retrieval (Soleimani et al., 2020).

Additionally, there is a pre-trained German language model available "bert-base-german-cased" (Chan et al., 2020) that was trained, among other sources, on a large corpus of legal texts. It is trained on cased German texts and, like the original BERT model, has 12 hidden layers with a size of 768, 12 attention heads per attention layer, and 110 million parameters. The model was trained on the German Wikipedia and a web corpus gathered by Suárez et al. (2019), which account for more than 90% of the data the model was trained on. However, the model was also trained on the Open Legal Data set from Ostendorff et al. (2020), which consists of more than 100,000 German court decisions.

6 Evaluation

We used the HuggingFace transformers library (Wolf et al., 2019) to fine-tune the pre-trained language model with our data set on the binary classification task of deciding whether a clause is void or not. We split our corpus into a training (80%) and a test set (20%) and first perform a stratified five-fold cross-validation on the training set to identify the best performing hyper-parameters for the fine-tuning. We started our search with the values suggested in the original BERT paper: batch size 16 or 32, learning rate 5e-5, 3e-5 or 2e-5, and 2, 3 or 4 epochs (Devlin et al., 2019). However, the authors also note that the optimal hyper-parameters are task-specific and that small data sets (which they define as less than 100,000 labels) are more sensitive to the choice of parameters than larger ones, therefore we also tried a smaller batch size (8) and higher numbers of epochs (8, 12, 16, 21). In the end, we found that batch size 16, learning rate 3e-5, and three epochs performed best.

With these hyper-parameters, we evaluated the approach on our test data set, which consists of 237 clauses, of which 192 are valid and 45 are void. BERT performed very well in the classification of void clauses and achieved an accuracy of 0.9, as well as a precision and recall of 0.9.

Out of the 45 void clauses in the test data, only four clauses have falsely not been identified as void (false negatives). Since our approach is meant to be a support tool for experts, all results will be double-checked by a human expert, which makes a high recall desirable.

A deeper analysis of the results showed that, while some types of clauses, as mentioned before, are virtually always void in the data set, others are virtually never. This might have (positively) influenced the classification performance.

7 Ethical and Societal Implications

The goal of this work is to support consumer advocates in order to further consumer protection and address the imbalance of power between corporations and consumers. While these are, by most standards, worthy and ethical goals, just because something is well-intended does not mean it can not have critical or at least ambivalent consequences. In this section, we want to highlight some of the issues that can arise from the research presented in this thesis and the goals it pursues. The laws governing T&C are changing comparably fast. For small

companies, without in-house legal counselling, it can therefore be expensive and challenging to keep up with the changing legislation and keep T&C always up to date. In such cases, honest mistakes might be made in drafting and maintaining T&C which do not intend to harm consumers. Nevertheless, such mistakes can make companies vulnerable to cease-and-desist orders from competitors and organisations which specialise in sending out cease-and-desist orders, not in order to protect consumer interests but for personal financial benefit. Therefore, we choose organisations to collaborate with that are dedicated to consumer protection and bound to that aim by their statute and their state given mission. However, it can not be prevented that our research can also be used by less well-intended actors. While this poses a potential threat, it can also allow companies on the other side to use our results in the same way on their own T&C and hence make sure they match the rule of law.

A second, arguably more philosophical issue that arises, not just from our research, but from the perspective of consumer-focused LegalTech in general, is whether our legal system is prepared for lowering the bar for accessing the system. The legal and moral standpoint on this issue is quite clear. The charter of fundamental rights of the EU guarantees in article 47 that "everyone whose rights and freedoms guaranteed by the law of the Union are violated has the right to an effective remedy before a tribunal". While the legal situation is clear, it is also clear that there are, in fact, barriers in place which make access to justice harder, whether they are of financial or procedural nature. And while it could be denied that they purposefully do so, it is difficult to deny that these barriers help to keep up the in many countries already stretched legal systems. If we would be able to denounce our neighbours by the click of a button every time they disturb the nighttime, this could not just have implications for the viability of our legal systems but also for the kind of society we live in and how we interact with each other. Concerning our work, we would argue that, if it has any influence on the legal system at all, it is designed to reduce its load. While the number of cease-and-desist orders sent out by consumer advocates might rise, we would hope that subsequently, this would lead to fewer cases brought on by consumers about void clauses in T&C.

Finally, if a system that automatically T&C for their lawfulness would be successful and widely adopted, one of the implications would very likely be that companies could start trying to "gamble" the system. This is a phenomenon that can be observed very well in the area of search engine optimisation (Malaga, 2008) and security (Mansfield-Devine, 2018). This could potentially lead to a situation where such a system would mostly fail to detect clauses that were purposefully drafted in a consumer-aversive way and would potentially be left detecting mostly clauses that are unintentionally void, e.g., by honest mistake, and were never intended to harm consumers. If we can learn anything from search engine optimisation and security, then that there is no easy or permanent fix to such problems. We, therefore, try to build our system in a way that it can be easily adapted, so that consumer advocates can change the system in a way that it will be able to detect such clauses, once they became aware of it, entering an "arms race" with malicious companies. And while "security through obscurity" is generally discouraged, search engine providers have shown that obfuscating the exact criteria helps to stay ahead of attempts to manipulate the ranking of websites. Therefore, our decision to focus on consumer advocates as users, rather than consumers themselves, can also help to mitigate the problem since companies will not be able to directly test different versions of their clauses.

8 Conclusion

In this paper, we have given an example of how NLP can be used to further the goal of consumer protection and address the existing imbalance of power between consumers and companies. We have argued that, in order to support consumers as broadly and effectively as possible, one should not (only) target individual consumers as potential users, but rather target organisations that represent consumers and their interests and have the power and means to pursue legal battles.

Together with experts from consumer protection agencies, we labelled a corpus of more than 1,100 German clauses from T&C from online shops with regard to their lawfulness. We showed that the labelling process already generated an impact on consumer protection, by enabling consumer advocates to send cease-and-desist orders against clauses that were identified as void and by providing new insights to consumer advocates, e.g. about the average share of void clauses in T&C.

We used this corpus to fine-tune a pre-trained BERT model that can identify void clauses in T%C with an accuracy of 0.9.

So far, the project and the developed classifier resulted in ten cease-and-desist orders that were sent to companies using void clauses in their T&C and hence protecting potentially hundreds of consumers. The approach is currently used in a test mode by two NGOs. By further integrating the technology into the existing workflows of consumer protection agencies and building a pipeline to continuously improving the model, based on manual annotations and corrections made by experts, we hope to be able to contribute to the protection of many more consumers in the future.

Acknowledgements

References

Reuben Binns and David Matthews. 2014. 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*, pages 881–884. ACM.

Daniel Braun and Florian Matthes. 2020. Automatic detection of terms and conditions in german and english online shops. In *Proceedings of the 16th International Conference on Web Information Systems and Technologies - WEBIST,*, pages 233–237. INSTICC, SciTePress.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. 2017. Satos: Assessing and summarising terms of services from german webshops. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 223–227, Santiago de Compostela, Spain. Association for Computational Linguistics.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. 2018. Customer-centered legaltech: Automated analysis of standard form contracts. In *Tagungsband Internationales Rechtsinformatik Symposium (IRIS) 2018*, pages 627–634. Editions Weblaw.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. 2019a. Consumer protection in the digital era: The potential of customer-centered legaltech. In *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft*, pages 407–420, Bonn. Gesellschaft für Informatik e.V.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. 2019b. The potential of customer-centered legaltech. *Datenschutz und Datensicherheit - DuD*, 43(12):760–766.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *arXiv preprint arXiv:2010.10906*.

Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans-W Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2018a. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. Technical report, European Consumer Organisation (BEUC).

Giuseppe Contissa, Francesca Lagioia, Marco Lippi, Hans-Wolfgang Micklitz, Przemyslaw Palka, Giovanni Sartor, and Paolo Torroni. 2018b. Towards consumer-empowering artificial intelligence. In *International Joint Conference on Artificial Intelligence*, pages 5150–5157.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Parliament and Council of the European Union. 2013. Regulation (eu) no 524/2013 of the european parliament and of the council of 21 may 2013 on online dispute resolution for consumer disputes and amending regulation (ec) no 2006/2004 and directive 2009/22/ec (regulation on consumer odr). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0524. Accessed 2020-07-10.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548.

Ruta Liepina, Giuseppe Contissa, Kasper Drazewski, Francesca Lagioia, Marco Lippi, Hans-Wolfgang

Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2019. Gdpr privacy policies in claudette: Challenges of omission, context and multilingualism. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 17th International Conference on Artificial Intelligence and Law (ICAIL 2019)*.

Marco Lippi, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Przemysław Pałka, Giovanni Sartor, and Paolo Torroni. 2019a. Consumer protection requires artificial intelligence. *Nature Machine Intelligence*, 1(4):168–169.

Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Yannis Panagis, Giovanni Sartor, and Paolo Torroni. 2017. Automated detection of unfair clauses in online consumer contracts. In *JURIX*, pages 145–154.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019b. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019c. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Ross A Malaga. 2008. Worst practices in search engine optimization. *Communications of the ACM*, 51(12):147–150.

Steve Mansfield-Devine. 2018. The malware arms race. *Computer Fraud & Security*, 2018(2):15–20.

Hans-W Micklitz, Przemysław Pałka, and Yannis Panagis. 2017. The empire strikes back: digital control of unfair terms of online services. *Journal of consumer policy*, 40(3):367–388.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an open platform for legal information. *arXiv preprint arXiv:2005.13342*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *European Conference on Information Retrieval*, pages 359–366. Springer.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Damiano Torre, Sallam Abualhaija, Mehrdad Sabetzadeh, Lionel Briand, Katrien Baetens, Peter Goes, and Sylvie Forastier. 2020. An ai-assisted approach for checking the completeness of privacy policies against gdpr. In *in Proceedings of the 28th IEEE International Requirements Engineering Conference (RE'20)*.

Verbraucherzentrale Bundesverband e.V. 2019. Jahresbericht 2018. https://www.vzbv.de/content/bericht-2018. Accessed 2020-05-04.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.