Automatic Detection of Terms and Conditions in German and English Online Shops

Daniel Braun[a] and Florian Matthes[b]
Department of Informatics, Technical University of Munich, Boltzmannstrasse 3, Garching, Germany

Keywords: Natural Language Processing, Document Classification, Terms and Conditions, Web Crawling.

Abstract: Terms and Conditions in online shops are arguably among the most important (or at least the most widely used) forms of consumer contracts. At the same time, they are probably among the least read documents. Thus, their automated analysis is of great interest, not just for research, but also from a consumer protection perspective. To be able to automatically process large amounts of Terms and Conditions and build the corpora which are necessary to train data-driven systems, we need means to identify Terms and Conditions automatically. In this paper, we present and evaluate four different approaches to the automatic detection of Terms and Conditions pages in German and English online shops. We treat the problem as a binary document classification problem for web-pages and report an approach which achieves precision, recall, and F1-score above 0.9 in German and close to 0.9 in English, by analysing the URL of the page.

1 INTRODUCTION

As consumers and internet users, we are confronted with Terms and Conditions (T&C) on a daily basis. Their content governs what happens to our private data, for how long we can return goods we do not like, who pays for their shipping, and many other things. While they are arguably among the most important consumer contracts nowadays, they are hardly ever read. Even if they are read, they are often so difficult to understand, that consumers still do not understand what they agree to (Bakos et al., 2014; Obar and Oeldorf-Hirsch, 2020). Services like "ToS; DR" (Binns and Matthews, 2014) offer crowd-sourced summarizations Terms of Services (ToS) and T&C in order to support consumers in making informed decisions. However, they only cover a limited number of popular websites and are sometimes based on outdated versions of their T&C.

In recent years, researchers from different institutions, like Braun et al. (Braun et al., 2017, 2018, 2019b,a) and Lippi et al. (Lippi et al., 2017; Micklitz et al., 2017; Lippi et al., 2019), have investigated techniques to automate the detection of unfair and illegal clauses in T&C. To scale that detection across as many online shops as possible, but also to gather data to train and test such approaches, the detection of T&C pages has to be automated. In this paper, we present four different approaches to this task for German and English. From a Natural Language Processing (NLP) perspective, the problem at hand is as a binary document classification task.

2 RELATED WORK

Document classification in the legal domain is widely used in the so-called "e-discovery" process, to find privileged or relevant documents in large corpora of various document types. E-discovery in this context is "any process (or series of processes) in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case" (Conrad, 2010). There is a vast amount of literature available on the use of document retrieval and classification to automate or assist the e-discovery process, including Brown (2015); Akers et al. (2011); Ayetiran (2013); Conrad (2010) and Mauritz (2018), as well commercial tools.

Other legal document classification tasks include the labelling of EU documents based on the EuroVoc thesaurus, as presented by Hajlaoui et al. (2014); Elnaggar et al. (2018) and Chalkidis et al. (2019), and the labelling of court decisions with the respective area of law they fall into (Soh et al., 2019).

[a] https://orcid.org/0000-0001-8120-3368
[b] https://orcid.org/0000-0002-6667-5452

Braun, D. and Matthes, F.
Automatic Detection of Terms and Conditions in German and English Online Shops.
DOI: 10.5220/0010154302330237
In *Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020)*, pages 233-237
ISBN: 978-989-758-478-7

The only previous work directly addressing the automatic detection of T&C, by Braun et al. (2017), presented two approaches for the automatic detection of T&C in German online shops: a rule-based URL classifier and a naive Bayes text classifier. Both were trained and tested on a set of 3424 URLs, which included 2592 T&C pages, extracted from Idealo, and 832 others pages from online shops. The URL classifier achieved an F1-score of 0.64 and the text classifier an F1-score of 0.86. Table 2 shows that even our worst-performing approach still performs better and our URL classifier even achieves an F1-score of 0.95 and thus improves the state of the art significantly.

3 CORPUS

To test and train the different classification approaches, we needed a sufficiently large corpus of labelled T&C from German and English online shops, but also of other, non-T&C pages. Since, to the best of our knowledge, no such corpus is currently available, we built our own corpus by scraping the list of merchants from two German price comparison websites ("Idealo"[1] and "Geizhals"[2]) that also offer versions of their website targeted to the British market and British online shops[3].

On both websites, shop operators manually report the URLs to their T&C pages. In this way, we could crawl 4,459 manually annotated links to German T&C from Idealo and 1,335 from Geizhals. After removing duplicates, our corpus consisted of 4,875 distinct links. We were able to download 4,869 pages from this 4,875 links. Five could not be downloaded because the websites were (permanently or temporarily) not available. In order to also have negative samples in our corpus (i.e. pages that do not contain T&C), we also downloaded the landing page of each shop (4,852) and a random other non-landing and non-T&C page (4,687). We chose this page by randomly selecting an outgoing internal link from the landing page and checking that it does not link to the T&C page or the landing page itself.

We performed the same process for the English versions of both websites and were able to download 543 T&C pages, 549 landing pages, and 486 random other pages. Table 1 shows an overview of the corpus. In total, our corpus consists of 15,986 HTML pages and their URLs.

We used the Article Extractor from BoilerPy3[4] to extract the plain T&C text from the HTML pages, by removing HTML tags, but also other noise, like navigation menus and footers. Since a surprisingly high number of pages was comprised of invalid HTML, which can not be processed by BoilerPy, we first ran all pages through Beautiful Soup[5] to fix the invalid HTML markup.

Table 1: Corpus of pages from German and English online shops.

Language	Page Type	Number of Pages
German	T&C	4,869
	Landing	4,852
	Other	4,687
	Σ	*14,408*
English	T&C	543
	Landing	549
	Other	486
	Σ	*1,578*
Total	Σ	*15,986*

While we cannot publish the corpus of HTML files due to copyright reasons, the URLs of all pages that are part of the corpus are available under a Creative Commons license on GitHub: https://github.com/sebischair/TC-Detection-Corpus. Please be aware that by the time of publication, the content of the pages might have changed, and some URLs might not be valid anymore.

4 CLASSIFICATION APPROACHES AND EVALUATION

In this section, we compare four different approaches to automatically find T&C pages in German and English online shops, by performing a binary document classification. We compare two rule-based and two stochastic approaches. While the stochastic approaches work with the content of the web pages, the rule-based approaches consider meta-information about the documents. Since we work with web pages, in addition to the textual content, we take into account the URL of each document and the text of the links that refer to the page as additional features for the classification.

4.1 Rule-based URL Analysis

The first approach we investigated is based on the URL of a web page. It splits the URL into its components and then performs a keyword search on them.

[1] www.idealo.de
[2] www.geizhals.de
[3] www.idealo.co.uk and www.skinflint.co.uk
[4] https://github.com/jmriebold/BoilerPy3
[5] https://www.crummy.com/software/BeautifulSoup/

Table 2: Evaluation of different techniques for the automatic detection of T&C in German and English online shops (best results for each metric and language are highlighted).

Language	Technique	Precision	Recall	F1-score
German	Rule-based URL analysis	0.99	0.91	0.95
	Logistic regression	0.98	0.82	0.89
	Neural network	0.96	0.83	0.88
	Neural network (multiling.)	0.95	0.84	0.89
English	Rule-based URL analysis	0.98	0.81	0.88
	Logistic regression	0.96	0.72	0.81
	Neural network	0.90	0.72	0.80
	Neural network (multiling.)	0.95	0.72	0.82

The keywords we use are “agb”, “geschaeftsbedingungen”, “geschaftsbedingungen”, “terms”, “conditions”, “gtc”, “tcs”, and “tac”. Since we do not need any training data for this rule-based approach, we could evaluate it on the whole corpus, i.e. 14,408 URLs from German online shops and 1,578 URLs from English online shops.

As shown in Table 2, the approach performed very well on both languages, with a precision of 0.99 for German and 0.98 for English and a recall of 0.91 and 0.81 respectively. Most false negatives i.e. missed T&C pages, originated from URLs with generic, mostly numerical, IDs. Some URLs in the German corpus contain misspelt versions of “geschaeftsbedingungen” and could therefore not be detected by this approach. One reason for the slightly lower precision that was achieved on the English corpus is the fact that there were more false positives because some English words contain the letter sequence “tac”, which is one of our keywords. Most notably the word “contact”. Although more complex rules could improve the precision, we found this simplistic approach as it is a very strong baseline for both languages.

4.2 Rule-based Link-text Analysis

Secondly, we used a rule-based approach that is not based on the URL of a link, but the link text(i.e. the text between the `<a></a>` tags). This approach can not be evaluated in the same fashion as the other approaches since it does not take a link or its content as input, but analyses all outgoing links from a page. Therefore, we evaluated the approach by parsing all landing pages from both corpora and checking whether the T&C link that is also in the corpus could be retrieved. The approach is also based on a relatively simple keyword search. The list of keywords this time was: “agb”, “allgemeine geschäftsbedingungen”, “geschaftsbedingungen”, “terms”, “conditions”, and “t&c”. From 4,837 German landing pages, we could find the correct T&C page with this method in 4,383 cases. In 454 cases the T&C page was not found. In no case, a wrong page was identified as T&C. Hence, we were able to extract the T&C page correctly in 90.6% of all cases. On the English corpus, we were able to extract the T&C page from 425 of 539 landing pages, which is 78.8% of all cases.

The other cases, in which we were not able to find the T&C page, do not necessarily represent a misclassification of the algorithm, since it is not mandatory for online shops to link to their T&C from their landing page. Therefore, there might be cases in which the approach correctly did not extract any link. Some spot tests in the English corpus indeed show that such cases exist. The percentage value of correctly extracted links can, therefore, alone, not be used as a measurement of correctness. However, the value does give an idea of the usefulness of the technique in practice.

4.3 Logistic Regression

The first stochastic approach we evaluated was logistic regression. We used Scikit-learn Pedregosa et al. (2011) to train a logistic regression classifier on a bag-of-words model of the corpora. We split the corpus into a training and a test data-set (80%/20%) while maintaining the distribution of both classes from the original corpus (roughly 2:1). The result was a German training data-set with approximately 11,500 documents and an English training data-set with approximately 1,200 documents. We used the training data to perform a grid search with a stratified 10-fold cross-validation, to find the parameter for the regularisation strength. We found that the parameter $C = 0.01$ performed best for both languages.

Table 2 shows that logistic regression worked well in both languages (F1-score 0.89 in German and 0.81 in English), despite the relatively sparse training data in English. However, it did not perform as good as the rule-based URL analysis.

4.4 Neural Network

In a final experiment, we used document embeddings to train a deep neural network. We used a pre-trained model from Google, called the Multilingual Universal Sentence Encoder model[6] (Yang et al., 2019), to generate the document embeddings. The model is trained on 16 languages, including German and English. We used the document embeddings to train a neural network, using Keras[7].

The network has one hidden layer that uses the ReLU activation function. A dropout layer was added to prevent overfitting and an output layer with one neuron that uses the sigmoid activation function. We again performed a grid search to find the best parameters for the batch size, the number of epochs, the dropout rate, and the number of neurons in the hidden layer. We ended up using a batch size of 20, 200 epochs, a dropout rate of 0.1, and 110 neurons.

We used the same training/test split as before. First, we trained and tested only on the same language, in a second step, we also tried to train the classifier jointly on both languages. The results of the evaluation are shown in Table 2. For both languages, the neural network performed similary to the logistic regression (F1-score 0.88 in German and 0.80 in English). The rule-based URL analysis approach, however, still performed significantly better in both cases. In both languages, the results could be slightly improved by training jointly on both languages.

5 INTERPRETATION

The rule-based URL analysis performed exceptionally well and achieved a precision close to 1.0 for German and English and also achieved the best precision, recall, and F1-score of all approaches on the German corpus. In addition to generating excellent classifications, the approach is also very fast and can classify the whole corpus within seconds. On the English corpus, the rule-based approach also performed best with regard to all three metrics.

An advantage of the neural network in comparison to the URL analyser is the fact, that the neural network uses the actual text for the classification. This approach is, therefore, better transferable to other applications and could, e.g. also be used to classify emails or PDF documents. Moreover, since we use multilingual embeddings, we are also able to transfer our models between different languages and still receive useful results.

[6]https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3

[7]https://keras.io/

6 CONCLUSION

We presented and compared four different approaches to the automatic detection of T&C in German and English online shops. Two approaches are based on the textual content (logistic regression with a bag of words model and a neural network with document embeddings) and two are based on meta-information (rule-based analysis of link URLs and texts). The results show that the rule-based URL analysis performed best in both languages with regard to precision, recall and F1-score. The results of the neural network could in both languages be slightly improved by training the network jointly on both languages.

In the future, we would like to investigate whether the models we trained could be transferred to

- (a) different languages: the universal sentence encoder model we used to calculate the sentence embeddings supports 14 more languages
- (b) other types of documents: we would like to investigate whether the model we trained for the deep learning approach could also be applied to related forms of standard form contracts, like insurance conditions or tenancy agreements.

ACKNOWLEDGEMENTS

The project is supported by funds of the Federal Ministry of Justice and Consumer Protection (BMJV) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme.

REFERENCES

Akers, S., Mason, J. K., and Mansmann, P. L. (2011). An intelligent approach to e-discovery. In *DESI IV: The ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*.

Ayetiran, E. F. (2013). An intelligent hybrid approach for improving recall in electronic discovery. In *DoCoPe@ JURIX*.

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

Binns, R. and Matthews, D. (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.

Braun, D., Scepankova, E., Holl, P., and Matthes, F. (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.

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

Braun, D., Scepankova, E., Holl, P., and Matthes, F. (2019a). Consumer protection in the digital era: The potential of customer-centered legaltech. In David, K., Geihs, K., Lange, M., and Stumme, G., editors, *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft*, pages 407–420, Bonn. Gesellschaft für Informatik e.V.

Braun, D., Scepankova, E., Holl, P., and Matthes, F. (2019b). The potential of customer-centered legaltech. *Datenschutz und Datensicherheit - DuD*, 43(12):760–766.

Brown, S. (2015). Peeking inside the black box: A preliminary survey of technology assisted review (tar) and predictive coding algorithms for ediscovery. *Suffolk J. Trial & App. Advoc.*, 21:221.

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2019). Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Conrad, J. G. (2010). E-discovery revisited: the need for artificial intelligence beyond information retrieval. *Artificial Intelligence and Law*, 18(4):321–345.

Elnaggar, A., Gebendorfer, C., Glaser, I., and Matthes, F. (2018). Multi-task deep learning for legal document translation, summarization and multi-label classification. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, pages 9–15.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). Dcep-digital corpus of the european parliament. In *LREC*, pages 3164–3171.

Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Panagis, Y., Sartor, G., and Torroni, P. (2017). Automated detection of unfair clauses in online consumer contracts. In *JURIX*, pages 145–154.

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., and Torroni, P. (2019). Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Mauritz, B. J. (2018). Automatic classification of legal documents. Master's thesis, Masarykova univerzita.

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

Obar, J. A. and Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Soh, J., Lim, H. K., and Chai, I. E. (2019). Legal area classification: A comparative study of text classifiers on singapore supreme court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2019). Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.