Clause Topic Classification in German and English Standard Form Contracts

Daniel Braun
University of Twente
Department of High-tech Business and Entrepreneurship
d.braun@utwente.nl

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

Abstract

So-called standard form contracts, i.e. contracts that are drafted unilaterally by one party, like terms and conditions of online shops or terms of services of social networks, are cornerstones of our modern economy. Their processing is, therefore, of significant practical value. Often, the sheer size of these contracts allows the drafting party to hide unfavourable terms from the other party. In this paper, we compare different approaches for automatically classifying the topics of clauses in standard form contracts, based on a data-set of more than 6,000 clauses from more than 170 contracts, which we collected from German and English online shops and annotated based on a taxonomy of clause topics, that we developed together with legal experts. We will show that, in our comparison of seven approaches, from simple keyword matching to transformer language models, BERT performed best with an F1-score of up to 0.91, however much simpler and computationally cheaper models like logistic regression also achieved similarly good results of up to 0.87.

1 Introduction

So-called standard form contracts, i.e. contracts that are drafted unilaterally by one party of the contract, usually a company, like terms and conditions of online shops or terms of services of social networks, are cornerstones of our modern economy. While the concept of a contract that is completely decided upon by one party might seem unfair and inherently flawed, its existence is a necessity in our modern economy. It would simply not be possible for companies like Amazon, Facebook or Google, to negotiate individual contract terms with each of their customers.

It is largely acknowledged, that most consumers do not read such contracts before buying something online or registering for a service. The actual share of consumers that regularly read such contracts ranges from as little as 3.5% (Plaut and Bartlett III, 2012) to 9% (Braun, 2021) in the literature. In acknowledgement of this fact, lawmakers around the world have tightly restricted the provisions that can be made by standard form contracts in a bid to protect consumers. This makes them an interesting subject for different Natural Language Processing (NLP) tasks, because they have a high economical and therefore practical relevance, but are still somewhat restricted with regard to their content. In this work, we use contracts that have been drafted under the jurisdiction of the European Union (EU). Many of the regulations applying to standard form contracts in this jurisdiction originate from the Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts.

While standard form contracts are relevant in almost all kinds of business-to-consumer transactions, like banking, insurances, and data processing, we here focus on Terms and Conditions (T&C) from online shops for three reasons: They have high economical relevance, they are publicly available in large quantities on the internet in a machine-readable format, and they are among the contracts which are least likely to be read, compared to a contract for life insurance, for example.

Being able to automatically classify the topics of clauses in T&C could help consumers to find relevant regulations faster and make more informed decisions, but it could also support legal professionals, like lawyers specialising in consumer protection law, in their work.

2 Related Work

Contract review is one of the main commercial applications of NLP in the legal domain (Dale, 2019). Unlike standard form contracts, "normal" contracts, i.e. contracts that have been negotiated by all contracting parties, allow for more variation and are less regulated, especially in business-to-business contexts. Therefore, building a taxonomy of clause

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 199 - 209

topics and performing topic classification (i.e. a supervised approach), would be less suitable for such contracts.

For reasons of data availability, most research projects use publicly available contracts, which happen to be standard form contracts, like T&C from online shops, Terms of Services (ToS) from online platforms, and, since the introduction of the General Data Protection Regulation (GDPR) in the EU, one particular focus has been on privacy policies. Although, legally, it is not settled whether they have a contractual status or not (Raysman and Brown, 2010), from an NLP perspective, they can be treated as a special variety of standard form contracts.

Most of the existing research on the analysis of standard form contracts is focusing on automatically finding void clauses. The CLAUDETTE project (Lippi et al., 2019), for example, focuses on finding so-called "unfair clauses", which are void under EU legislation, in ToS from large online platforms like Facebook or Netflix. Later, they also applied their approach to privacy policies (Liepina et al., 2019). Similarly, Braun and Matthes (2021) focus on finding void clauses in T&C from online shops by using a fine-tuned BERT model, in earlier work, they summarised specific aspects of T&C using an abstractive summarisation approach (Braun et al., 2017). In comparison to these works, which try to make a fully automated decision on the validity of clauses, the classification of clause topics could be used to support humans in the decision-making process, by helping them to find relevant clauses faster.

In the area of privacy policies, approaches are more diverse. Ravichander et al. (2019), for example, presented a Q&A system that can answer user questions about privacy policies. Binary assessments in classes like valid or void are less desirable in the domain of privacy policies where, especially before the introduction of the GDPR, much of what was legally allowed was still undesired by users.

3 Taxonomy

For a topic classification approach, i.e., a supervised approach, rather than an unsupervised topic modelling approach, a taxonomy of clause topics is needed. At first, this might seem like a limitation of the approach, because, in theory, a contract can regulate arbitrary aspects. In practice, however, standard form contracts underly strong limitations, because, under EU jurisdiction, clauses that are "unexpected" are automatically void under the Unfair Contract Terms Directive (93/13/EEC). Therefore, if a taxonomy is extensive enough, the information that a clause is not covered by one of the topics in the taxonomy is already important information in itself, because it means that the clause is very likely void.

To build such an extensive taxonomy for T&C from online shops, we used contract templates from legal literature (Sommer and von Stumm, 2017; Fingerhut, 2009) and industry associations (IHK Munich and Upper Bavaria, 2020; Schirmbacher, 2018), as well as a commercial T&C generator[1] and analysed which topics are present in these templates, because they are used by many online shops.

For each of these sources, two legal experts with experience in consumer protection law went through the templates and annotated each clause with one or more fitting topic and zero or more fitting subtopic label(s). In the end, the topic labels from the different annotators were aligned by the authors. By the combination of the above-described sources, we derived a taxonomy of 22 classes (topics) and 36 sub-classes (subtopics). The differentiation between topics and subtopics was mainly based on the structure of the sources, i.e. the organisation in sections and subsections in the templates. A topic, for example, could be "delivery" and subtopics could be the delivery time or the delivery costs.

To get an estimate of how extensive our taxonomy is, we used it to manually annotate more than 6.000 clauses from real T&C (see section 4). Of these more than 6,000 clauses, the taxonomy was able to cover 90.08%. The remaining clauses (336) fell into only two classes: 285 clauses contained information regarding vouchers and gift cards, and 51 clauses contained information about codes of conduct. We added both classes to the taxonomy. The final taxonomy, therefore, consists of 23 labels for topics and 37 labels for subtopics. All labels in the taxonomy are shown in Table 2.

4 Corpus

Since no corpus of topic-annotated clauses from T&C existed, we had to build our own corpus. For this, we parsed the list of merchants from two German price comparison websites ("Idealo"[2] and

[1] https://www.trustedshops.com
[2] www.idealo.de

"Geizhals"[3]) that also offer a localised version of their respective websites in English, targeted to the British market[4]. On these websites, shop operators manually report the URLs to their T&C, which we extracted with a web-crawler. We randomly selected 142 German T&C and 30 English T&C from these pages. Each clause from these contracts was subsequently copied into an Excel file, in which each row contains one clause. In addition to the text of the clause itself, each row contains a unique id, an id for the contract the clause belongs to, (if existing) the title of the superordinate paragraph and (if existing) the title of the clause.

4.1 Size

The corpus we built consists of 5,020 German clauses and 1,040 English clauses. In both languages, a contract, therefore, consists of roughly 35 clauses on average. All German clauses together consist of 351,903 words, which is an average of 2,478 words per contract (see Table 1). The English corpus contains 55,392 words which equals to an average of 1,846 words per contract. This means German clauses are, on average, significantly longer than English ones. 5,013 clauses (or 99.9% of all clauses) in the German corpus have a paragraph or clause title (or both), which we can use for the topic classification. In the English corpus, that is the case for 989 clauses or 95.1%.

4.2 Annotation

Each clause of both corpora was labelled with its topics and subtopics according to the taxonomy described in Section 3. First, each clause was labelled by a student using only the classes from the first level (topics) of the taxonomy. Then, where applicable, classes from the second level of the taxonomy (subtopics) were added. In a second step, this process was repeated by the authors, i.e., each clause was again first labelled with classes from the first level and then, where applicable, with classes from the second level. A clause can be labelled with more than one topic and subtopic. A clause can also be assigned to a topic without necessarily having to be assigned to a subtopic of it (but not the other way round). An example of such a clause from the corpus is "The warranty is subject to the relevant statutory provisions.", which is assigned the topic warranty, but not to one of its subtopics. Cases where the two annotators disagreed, were presented to (and finally decided by) consumer protection lawyers with many years of experience in advising consumers. The inter-annotator agreement was relatively high at 87%, i.e., only 13% of all clauses had to be decided by the lawyers. In this way, four people together spend more than 100 hours and generated more than 24,000 labels, which were consolidated into two corpora, one for each language, with 11,777 labels in total. The distribution of topics and subtopics is shown in Table 2.

The annotation also revealed local differences, e.g., almost none of the English contracts contained a model withdrawal form, the only two that did contain such a form were from companies based in Germany, while many German contracts contained one. On the other hand, clauses about loyalty schemes were almost non-existing in German contracts and far more popular in the English corpus. It is worth reminding that our English data set was collected specifically from a UK perspective, i.e., the shops are either based in the UK or specifically targeted at the UK market. English contracts from other markets, like the USA or Australia, would most likely look very different. Since the T&C we annotated are protected by copyright law, we are, unfortunately, not able to publish the corpus.

5 Approaches

We compared seven different approaches to the classification of clause topics in standard form contracts from German and English online shops: Rule-based keyword matching, Logistic Regression, Random Forest, Multilayer Perceptron (MLP), Long short-term memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT). In the following sections, we will shortly introduce how we used the different approaches. For each of the approaches, we trained two classifiers, one which only classifies topics and one which only classifies the subtopics. Experiments we conducted with joint classification models, i.e., models that classify topics and subtopics at the same time turned out to decrease the classification quality for both, topics and subtopics.

We split both corpora into a training (80%) and a test (20%) set, using scikit-multilearn (Szymański and Kajdanowicz, 2017) to make sure the representation of labels is balanced between the training and the test set and reflects the original distribution.

[3] www.geizhals.de

[4] www.idealo.co.uk, www.skinflint.co.uk

	contracts	clauses	words	∅ clauses/contract	∅ words/contract
German	142	5,020	351,903	35	2,478
English	30	1,040	1,846	34	1,846

Table 1: Statistics on the German and English corpus

For the stochastic approaches, we performed a grid search with a k-fold cross-validation on the training data to find the optimal parameters for each approach.

Our initial hypothesis, based on similar research, was, that with increasing complexity of the models, the performance would also increase, i.e. we expected BERT to perform best, followed by LSTM, MLP, and the "classic" ML approaches.

5.1 Rule-based

As a baseline, we first developed a rule-based classification approach. We used a simple keyword-matching approach. For each topic and subtopic in the taxonomy, we asked the consumer protection lawyers to provide a list of keywords that are distinctive for the topic/subtopic. The list can contain independent keywords (OR), keywords that should appear together (AND), and keywords that should not appear (together) (NOT).

We pre-processed the clauses using SoMaJo (Proisl and Uhrig, 2016) to split the clauses into sentences and the sentences into tokens. Afterwards, we lemmatised all tokens using the Stanford Lemmatizer (Manning et al., 2014) for English and the Mate tools Lemmatizer (Björkelund et al., 2010) for German before applying the rules. In German, we noticed that lemmatisation (but also stemming) face big challenges, especially in the legal domain, when it comes to compound nouns, i.e., nouns that are combined to create new nouns, like "Vertragspartner" (contractual partner) is a combination of "Vertrag" (contract) and "Partner" (partner). Compound nouns can be inflected internally ("Vertragspartner"), and splitting them into their constituents is not trivial. A "Druckerzeugnis" (printed matter) could, for example, lexically speaking either be a "Druck-Erzeugnis" (print - matter) or a "Drucker-Zeugnis" (printer - certificate). While there are existing approaches on how to automatically split compound words into their respective parts (e.g., by Baroni et al. (2002), Koehn and Knight (2003), Daiber et al. (2015), Sugisaki and Tuggener (2018), and Weller-Di Marco (2017)), the problem is far from being trivial and is not yet addressed in our implementation.

5.2 Logistic Regression

Second, we trained a logistic regression classifier, which we implemented using Scikit-learn (Pedregosa et al., 2011). As input, we used a Tf–idf vector representation of the concatenation of clause text and titles. Before transforming the clauses into these vectors, we removed stopwords using "Stopwords ISO"[5]. Since logistic regression does not inherently support multi-label classification, we used a "one-vs-the-rest" approach. Instead of training one classifier, we train one classifier for each class, which performs a binary classification against all remaining classes and combined all results to decide which labels are predicted for a given input.

We grid search with a 10-fold cross-validation on the training data to find the best parameter for the regularisation strength for the classification of topics. We performed multiple iterations on both languages to narrow down the search space. In German, we achieved the best results with $C = 1,000$ and in English with $C = 45,000$. The values are rather high for both languages but especially for the smaller English data-set. Since C is the inverse of the regulator $(1/\lambda)$, a high value for C means a low value for λ and hence poses the risk of overfitting. We performed the same procedure for the classification of subtopics and found that $C = 100$ performed best in both languages, which is significantly lower and therefore less prone to overfitting.

5.3 Random Forest

Logistic regression is computationally efficient and generalises well, and is, therefore, a good baseline. However, its inability for "real" multi-label classification is a drawback in our use case. Decision trees do inherently support multi-label classification and also are inherently explainable. However, they are not as efficient as logistic regression and are more prone to overfitting. Instead of training just one decision tree, we use a random forest approach, where multiple independent randomised

[5] https://github.com/stopwords-iso/stopwords-iso

Label	DE	EN	Total
age	38	5	43
applicability	253	33	286
applicableLaw	137	23	160
arbitration	155	13	168
changes	13	12	25
codeOfConduct	55	1	56
conclusionOfContract (cOc)	800	146	946
cOc:binding	328	39	367
cOc:changeOfOrder	58	6	64
cOc:definition	103	4	107
cOc:restrictions	42	7	49
cOc:steps	256	58	314
cOc:withdrawal	95	20	115
delivery	839	164	1003
delivery:brokenPackaging	134	10	144
delivery:costs	247	57	304
delivery:customs	43	6	49
delivery:destination	96	16	112
delivery:methods	160	17	177
delivery:partial	32	5	37
delivery:time	143	41	184
description	86	30	116
disposal	51	16	67
intellectualProperty	45	24	69
language	124	11	135
liability	439	140	579
party	157	21	178
payment	898	112	1010
payment:fee	50	3	53
payment:late	48	1	49
payment:loyalty	7	22	29
payment:methods	435	53	488
payment:restraint	46	1	47
payment:vouchers	301	14	315
personalData	213	49	262
personalData:cookies	6	3	9
personalData:duration	8	1	9
personalData:information	48	12	60
personalData:reason	50	11	61
personalData:update	7	4	11
personalData:usage	57	16	73
placeOfJurisdiction	117	19	136
prices	158	56	214
prices:currency	17	13	30
prices:vat	119	24	143
retentionOfTitle	222	13	235
severability	42	12	54
textStorage	152	11	163
warranty	540	25	565
warranty:options	69	5	74
warranty:period	155	10	165
withdrawal	484	202	686
withdrawal:compensation	94	27	121
withdrawal:effects	97	12	109
withdrawal:exclusion	100	27	127
withdrawal:form	131	37	168
withdrawal:model	41	2	43
withdrawal:period	126	40	166
withdrawal:shippingCosts	118	43	161
withdrawal:shippingMethod	74	13	87
Total lvl 1	6018	1138	7156
Total lvl 2	3941	680	4621

Table 2: Distribution of topic and subtopic labels among the German (DE) and English (EN) corpus

decision trees are trained, and a majority vote is used for classification. As input, we again used Tf-idf vectors.

We again performed a grid search with stratified 10-fold cross-validation on the training data to find the best performing values for the parameters: number of estimators (i.e., the numbers of trees), the maximum depth of the trees, the minimum number of samples per internal node that is needed for a split, and the minimum number of samples per leaf. As usual, we performed several iterations to narrow down the search space before, in the final iteration, we found the following values to perform best. In German: number of estimators = 2,000, maximum depth = ∞, samples per node = 2, samples per leaf = 1 and in English: number of estimators = 1,000, maximum depth = 100, samples per node = 2, samples per leaf = 1.

5.4 Neural Networks

For the different approaches using neural networks, we also evaluated different input encodings, namely different kinds of word embeddings. To train domain-specific embeddings, we used a larger corpus than the one described in Section 4, because the data does not have to be annotated. We collected the corpus in the same way as the other corpus from the price comparison websites, however, it is more than 30-times bigger, consisting of 5,412 contracts, 4,869 in German and 543 in English.

We used the following embeddings:

- German
  - Word2Vec embeddings with 300 dimensions based on the German Wikipedia[6]
  - GloVe embeddings with 300 dimensions based on the German Wikipedia[7]
  - Word2Vec embeddings with 300 dimensions we trained from scratch on the above described corpus of T&C
- English
  - Word2Vec embeddings with 300 dimensions based on the Google News Corpus (Mikolov et al., 2013)
  - GloVe embeddings with 300 dimensions based on Wikipedia and Gigawords 5 (Pennington et al., 2014)

[6] https://gitlab.com/deepset-ai/open-source/word2vec-embeddings-de

[7] https://gitlab.com/deepset-ai/open-source/glove-embeddings-de

– Word2Vec embeddings with 300 dimensions we trained from scratch on the above described corpus of T&C

Training neural networks is computationally much more expensive than, e.g., logistic regression and at the same time depends on more parameters. To manage the increasing complexity, we changed our approach for the parameter optimisation by reducing the 10-fold cross-validation to a 5-fold cross-validation. Additionally, we fixed parameters that always performed best or almost best, independent from other parameters, as early as possible in order to reduce the search space and converge faster to a local optimum.

5.4.1 MLP

The hyper-parameters we optimised of the MLP were the number of layers, the number of neurons per layer, the dropout, the batch size and the number of epochs. The results of the hyper-parameter studies can be found in Appendix A.

5.4.2 LSTM

For the LSTM network we optimised the sequence length, the number of LSTM layers and the number of neurons in them, the number of dense layers and the number of neurons in them, the dropout, the batch size, and the number of epochs. The best performing parameters we found for the topic and subtopic classification can be found in Appendix A.

5.5 BERT

Finally, we evaluate an approach using a transformer model for the clause topic classification, more specifically, the BERT language model (Devlin et al., 2019). We used the HuggingFace transformers library (Wolf et al., 2019) to fine-tune the pre-trained language models and implement the classification.

For English, we used the "bert-base-uncased" pre-trained model, provided by the original authors Devlin et al. (2019). The model, which is trained on lower case English texts, has 12 hidden layers with a size of 768, 12 attention heads per attention layer, and 110 million parameters. For German, we used the "bert-base-german-cased" model from Chan et al. (2020). It is trained on cased German texts and, like the original model, has 12 hidden layers with a size of 768, 12 attention heads per attention layer, and 110 million parameters.

The original BERT language model was trained on the English Wikipedia and the BookCorpus by Zhu et al. (2015), which consists of 11,038 fiction books that are available for free on the internet. The German language model we are using was pre-trained on a more diverse set of sources, among which are 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. We also briefly evaluated a multilingual approach with the Multilingual Universal Sentence Encoder transformer model, which was used by Braun and Matthes (2020) for the multilingual automated detection of T&C, however, first tests on German and English were not promising, so we did not follow through on the approach.

We used our training data to fine-tune both language models, the English and the German, for the topic classification task. In order to find the best hyper-parameters, we split 20% off the training data as validation set. 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. For our data sets and task, we found a smaller batch size with a slightly higher number of epochs to work better than the suggested parameters in both languages. We found a batch size of eight and a learning rate of 5e-5 to perform best for the topic and subtopic classification in both languages. In German, eight epochs performed best for the topic classification and six for the subtopic classification. In English, six epochs for the topic classification and 21 epochs for the subtopic classification performed best. All other parameters were kept equal to the original pre-trained model.

6 Evaluation

Our baseline approach of using keywords achieved an F1-score of 0.78 in German for topic classification and 0.64 for subtopic classification. The performance in English was worse with an F1-score of 0.72 for topics and 0.46 for subtopics. We noticed that, in German, the list of keywords

Approach	A	P	R	F1
BERT	0.84	0.93	0.89	0.91
Log. Regression	0.77	0.95	0.80	0.87
Random Forest	0.73	0.97	0.72	0.83
MLP	0.75	0.89	0.74	0.81
LSTM	0.73	0.90	0.72	0.80
Rule-based	0.64	0.77	0.80	0.78

(a) German

Approach	A	P	R	F1
BERT	0.79	0.89	0.82	0.85
Log. Regression	0.71	0.88	0.73	0.80
LSTM	0.72	0.80	0.74	0.77
MLP	0.72	0.79	0.73	0.76
Rule-based	0.57	0.76	0.69	0.72
Random Forest	0.57	0.88	0.58	0.70

(b) English

Table 3: Best clause topic classification results for each approach, ordered by F1-score (A = accuracy, P = precision, R = recall, F1 = F1-score)

Approach	A	P	R	F1
BERT	0.79	0.89	0.83	0.86
Log. Regression	0.75	0.91	0.78	0.84
Random Forest	0.68	0.91	0.67	0.77
MLP	0.73	0.86	0.66	0.75
LSTM	0.69	0.85	0.63	0.72
Rule-based	0.47	0.74	0.56	0.64

(a) German

Approach	A	P	R	F1
BERT	0.68	0.79	0.68	0.73
MLP	0.67	0.76	0.66	0.71
LSTM	0.67	0.78	0.65	0.70
Log. Regression	0.54	0.80	0.59	0.68
Random Forest	0.44	0.85	0.43	0.57
Rule-based	0.28	0.39	0.49	0.43

(b) English

Table 4: Best clause subtopic classification results for each approach, ordered by F1-score (A = accuracy, P = precision, R = recall, F1 = F1-score)

mostly consisted of domain-specific compound nouns, like "Widerrufsrecht" (right of withdrawal) or "Gefahrenübergang" (transfer of risk), which are very distinctive for their respective topics and make the classification relatively easy.

This is also a possible explanation for why the logistic regression classifier, with Tf-idf vectors as input, performed so well on the German corpus. With an F1-score of 0.87 for topics and 0.84 for subtopics, it was only surpassed by the BERT model. All other approaches performed comparable to each other in German, with F1-scores between 0.8 and 0.83 for topic classification and 0.72 to 0.77 for the subtopic classification (see Table 3 and 4).

In English, the picture was less clear, with logistic regression still performing best for topic classification (F1-score 0.80) but being surpassed by the neural network approaches for subtopic classification. However, the clear overall winner, with the best performance on both languages and classification levels was BERT, which scored up to 0.91.

For the approaches we evaluated with different inputs, i.e. the MLP and LSTM, the values in Table 3 and 4 represent the best results achieved. There was no clear pattern visible of which word embedding model performs best. All of them achieved comparable results and no one performed best or worse in all settings.

7 Transferability

Due to their practical relevance and availability, we focused on T&C from online shops in this paper. However, there is no reason why the same technology could not be applied to other types of standard form contracts, e.g. from banks and insurances. At the same time, the taxonomy we developed and the models we trained have a component that is specific to online shopping, broken packaging, for example, is a topic, that is not relevant for banking.

To get an idea of how domain-specific the taxonomy and the models are, we annotated the T&C of three of the largest German banks (Commerzbank, Deutsche Bank, and Sparkassen) in the same way described in Section 4.2: first, a student, then authors annotated the all clauses of the contracts independently with their topics and subtopics, then conflicting labels were resolved by the team of experts. We used the taxonomy described in Section 3 for the classification, however, we added a class "n.a." to mark clauses that cover a topic that is not represented by any of the classes in the taxonomy.

The three contracts consist of 214 clauses, about 71 clauses per contract, and 13,681 words, an aver-

Topic	#clauses
applicability	3
applicableLaw	3
arbitration	2
changes	5
liability	9
n.a.	143
payment	14
placeOfJurisdiction	6
prices	1
withdrawal	27

Table 5: Topics of clauses in the general business conditions of banks

Topic	P	R	F1
applicability	0.60	1.00	0.75
applicableLaw	1.00	1.00	1.00
arbitration	1.00	1.00	1.00
changes	1.00	0.83	0.91
liability	0.41	1.00	0.58
payment	0.16	1.00	0.28
placeOfJurisdiction	0.00	0.00	0.00
prices	0.00	0.00	0.00
withdrawal	0.54	1.00	0.70
TOTAL	0.28	0.97	0.43

Table 6: BERT clause topic classification results on the banking corpus

age of 2,478 words per contract. The contracts from the online shops, in comparison, consisted of an average of 35 clauses per contract. The 214 clauses consist of 13,681 words, which equals 4,560 words per contract. The fact that the banking contracts contain much more clauses per contract already suggests, that our taxonomy will, most likely, not be able to cover all of them.

The annotation process confirmed this assumption. Of the 214 clauses, only 71 (or 33%) are concerned with a topic that is covered by our taxonomy (see Table 5). The other clauses are concerned with a wide range of banking specific topics, from deposit protection funds to banking confidentiality. This means that, even if we would correctly classify all the other clauses, we could never achieve a recall above 0.33. We can already conclude, that the taxonomy we developed can not simply be applied to other types of standard form contracts without adaption.

Since the taxonomy is still able to cover one-third of the banking contracts, we wanted to test how well the classifiers we trained would perform on this data set. Therefore, we took the best performing topic classifier, i.e., the BERT model, and applied it to the new corpus. The evaluation of the results is shown in Table 6. We can see that for some of the topics, e.g., applicability, applicableLaw, arbitration, and changes, the performance is very good, even tough our model has never seen this type of contract before.

8 Conclusion

In this paper, we compared different approaches to classify the topics of clauses in standard form contracts from online shops, based on a taxonomy of clause topics and subtopics we developed and a bilingual corpus of more than 6,000 clauses we gathered and annotated. Our evaluation showed, that our initial hypothesis, that the model performance would increase with complexity, did not hold.

While BERT did indeed perform best for both languages, the much simpler logistic regression approach showed the second-best performance. Considering the computational time and power that is needed to not only train the more complex model but also during inference of the labels, the simple logistic regression approach might for some practical application be the better choice.

We were surprised to find that multilingual approaches, using the German and English data together, did not seem to bring any improvements for this task, even though earlier work in the legal domain, e.g. by Niklaus et al. (2021) and Braun and Matthes (2020), have shown that multilingual models can improve performance. This aspect needs further investigation.

While the taxonomy we developed and the models we trained are domain-specific for eCommerce, first tests suggest that the approaches can be transferred to other types of standard form contracts and that event the models can partially be transferred to other domains, at least for more "technical" clauses concerning the contract itself. This could apply to all types of consumer standard form contract within a highly regulated domain, like insurances, housing, and employment, and is something we would like to investigate further in the future.

Acknowledgements

The project was supported by funds of the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) 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

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Predicting the components of german nominal compounds. In *ECAI 2002: 15th European Conference on Artificial Intelligence*, pages 470–474.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China. Coling 2010 Organizing Committee.

Daniel Braun. 2021. *Automated Semantic Analysis, Legal Assessment, and Summarization of Standard Form Contracts*. Dissertation, Technische Universität München, München.

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 and Florian Matthes. 2021. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 93–99, Online. Association for Computational Linguistics.

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.

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

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. 2015. Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28, Praha, Czechia. ÚFAL MFF UK.

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

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.

Michael Fingerhut. 2009. *7. Teil: Allgemeine Geschäftsbedingungen*, 12th edition edition. Carl Heymanns Verlag.

IHK Munich and Upper Bavaria. 2020. Allgemeine Geschäftsbedingungen für einen Webshop. https://www.ihk-muenchen.de/ihk/documents/Recht-Steuern/Vertragsrecht/AGB-Webshop_2020.docx. Last accessed 2020-07-09.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings*

of the Natural Legal Language Processing Workshop 2021, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Victoria C Plaut and Robert P Bartlett III. 2012. Blind consent? a social psychological investigation of nonreadership of click-through agreements. *Law and human behavior*, 36(4):293.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62, Berlin. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Richard Raysman and Peter Brown. 2010. Contractual nature of online policies remains unsettled. *New York Law Journal*, 10.

Martin Schirmbacher. 2018. Allgemeine geschäftsbedingungen (online-shop). `https://www.bevh.org/fileadmin/content/01_leistungen/rechtshilfen/muster-agb/muster-agb-internetshop-2018.pdf`. Last accessed 2020-07-10.

Barbara Sommer and Ferdinans von Stumm. 2017. Fernabsatz von waren und dienstleistungen. In Wolfgang Weitnauer and Tilman Mueller-Stöfen, editors, *Beck-sches Formularbuch IT-Recht*, 4 edition, chapter J, pages 715–761. C. H. Beck Verlag, Munich.

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.

Kyoko Sugisaki and Don Tuggener. 2018. German compound splitting using the compound productivity of morphemes. In *14th Conference on Natural Language Processing-KONVENS 2018*, pages 141–147. Austrian Academy of Sciences Press.

P. Szymański and T. Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Marion Weller-Di Marco. 2017. Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 161–166, Valencia, Spain. Association for Computational Linguistics.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendix: Model Parameters

See Table 7 to 10, activation function for all was `tanh` and optimiser `adam`.

Language	Input	Layers	Neurons	Dropout	Batch	Epochs
German	T&C	2	200, 200	0.4	100	500
	Word2Vec	3	200, 150, 200	0.4	100	300
	GloVe	1	110	0.3	200	200
English	T&C	2	200, 150	0.2	20	300
	Word2Vec	3	200, 150, 200	0.3	100	500
	GloVe	3	100, 200, 100	0.2	50	400

Table 7: Hyper-parameters used for the topic classification with the Multilayer Perceptron on different inputs

Language	Input	Layers	Neurons	Dropout	Batch	Epochs
German	T&C	1	150	0.3	500	400
	Word2Vec	3	100, 200, 100	0.4	300	500
	GloVe	1	80	0.3	300	500
English	T&C	1	150	0.3	500	400
	Word2Vec	3	100, 200, 100	0.4	150	300
	GloVe	3	200, 150, 200	0.3	150	400

Table 8: Hyper-parameters used for the subtopic classification with the Multilayer Perceptron on different Inputs

Lang.	Input	Sequ. Length	LSTM Layers	Neurons	Dense Layers	Neurons	Dropout	Batch	Epochs
DE	T&C	35	1	300	2	200, 50	0.3	300	30
	W2V	50	1	300	1	50	0.3	40	30
	GloVe	50	1	300	1	50	0.3	40	13
EN	T&C	100	1	300	2	200, 50	0.3	150	100
	W2V	40	1	45	0		0.6	10	50
	GloVe	65	1	200	1	65	0.7	15	100

Table 9: Hyper-parameters used for the topic classification with the LSTM on different inputs

Lang.	Input	Sequ. Length	LSTM Layers	Neurons	Dense Layers	Neurons	Dropout	Batch	Epochs
DE	T&C	35	1	250	1	50	0.4	15	20
	W2V	45	1	200	1	50	0.4	15	20
	GloVe	45	1	105	1	50	0.3	15	10
EN	T&C	45	1	200	1	50	0.3	20	15
	W2V	50	1	250	1	50	0.4	25	25
	GloVe	40	1	150	1	50	0.3	15	10

Table 10: Hyper-parameters used for the subtopic classification with the LSTM on different inputs