Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 480–490
Marseille, 20-25 June 2022

Claim Extraction and Law Matching for COVID-19-related Legislation

Niklas Dehio, Malte Ostendorff, Georg Rehm
DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
niklas.dehio@gmail.com, {malte.ostendorff, georg.rehm}@dfki.de

Abstract

To cope with the COVID-19 pandemic, many jurisdictions have introduced new or altered existing legislation. Even though these new rules are often communicated to the public in news articles, it remains challenging for laypersons to learn about what is currently allowed or forbidden since news articles typically do not reference underlying laws. We investigate an automated approach to extract legal claims from news articles and to match the claims with their corresponding applicable laws. We examine the feasibility of the two tasks concerning claims about COVID-19-related laws from Berlin, Germany. For both tasks, we create and make publicly available the data sets and report the results of initial experiments. We obtain promising results with Transformer-based models that achieve 46.7 F1 for claim extraction and 91.4 F1 for law matching, albeit with some conceptual limitations. Furthermore, we discuss challenges of current machine learning approaches for legal language processing and their ability for complex legal reasoning tasks.

Keywords: Natural Legal Language Processing, Text Matching, Claim Extraction, Legal Reasoning, COVID-19

1. Introduction

The COVID-19 pandemic has led to a flurry of activity by law-making bodies in many jurisdictions around the world, for instance in Germany (Siewert et al., 2020). The rapid spread of the contagious coronavirus since December 2019 demanded both the necessity to pass laws that deeply intervene into our constitutional rights and to do so in a timely manner. However, the unpredictable nature of the pandemic lead to a situation where those laws were altered again and again, often on short notice, based on new insights produced by the scientific process.

While new rules were often communicated to the public via newspapers or public statements, these often did not provide a reference to the actual law itself. Finding the corresponding law is a non-trivial task in itself, also due to the technical language used in laws – also called “legalese” (Marín, 2017). In addition, understanding the applicability of laws requires specific reasoning skills that legal professionals like lawyers are trained in, but laypersons usually are not. Having an automated system that finds the referenced laws would be a valuable contribution to counter disinformation and enable citizens to look up legal rules themselves.

In this paper, we discuss the challenges of legal reasoning in the context of machine learning and examine the feasibility of two tasks. First, the claim extraction task, where *claims* about COVID-19-related legislation are extracted from newspaper articles. Second, the law matching task, where such claims are matched with relevant subsections from the corresponding laws. For both tasks, we annotate and make available data sets based on German news articles and German legislation.[1] Moreover, we report and discuss the results of initial experiments based on Transformer language models and a strong TF-IDF baseline.

2. Background and Related Work

In this section, we introduce basic legal terminology and relevant related work about fact-checking, claim retrieval, and legal reasoning.

2.1. Legal Terminology

In this article, we will call the collections of legal articles an *act* (also called *bill*, or *statute*, in German: “Gesetz”, “Verordnung”). An act consists of *sections*, which start with a “§” (also called “article”, or “clause”, in German called “Paragraph”). A section might be subdivided into several *subsections*, each start with an enclosed number[2].

2.2. Fact-Checking

Hanselowski et al. (2019) provide an overview of many popular fact checking data sets. Most of the work is domain-specific, e. g., about politics (Vlachos and Riedel, 2014; Wang, 2017; Atanasova et al., 2018), news (Ferreira and Vlachos, 2016), science (Wadden et al., 2020) or the web (Thorne et al., 2018; Gorrell et al., 2019).

2.3. Claim Retrieval

There is a rich body of research on claim detection. Pankovska et al. (2022) and Levy et al. (2014) investigate claim detection in context of COVID-19 and other specific topics. Beltran and Larraz (2021) present an automated claim detection tool for Twitter, that produces an 80% F1 score in real-life scenarios. Daxenberger et al. (2017) investigate claims across several data sets. Chakrabarty et al. (2019) train a claim detection model on a data set of 5.5 million claims from

[1] https://github.com/DFKI-NLP/covid19-law-matching

[2] The different wording for the same structure arise from differences between law systems, e. g., from the U.S. and U.K. In addition, there is no “official” translation for the German legal terms.

Reddit, leveraging transfer learning to achieve good results across several domains.

2.4. Natural Legal Language Processing

NLP in the legal domain is an emergent field. Chalkidis et al. (2021) present a benchmark and a collection of data sets to test language model performance across various legal tasks. Chalkidis et al. (2020) pre-train a BERT model on legal documents and report improved performance on legal related tasks. Ostendorff et al. (2021) evaluate document representations for legal literature recommendations. Schneider et al. (2021) develop a knowledge graph for the legal domain. Zhong et al. (2020b) present a multiple-choice question data set from the National Judicial Examination in China that requires advanced reasoning skills to solve. The Competition on Legal Information Extraction/Entailment workshop consists of five tasks concerning retrieval and inference over cases and statutes (Rabelo et al., 2021). Holzenberger and Van Durme (2021) model the task of determining whether a legal statute applies to a set of facts as four subtasks. Bommasani et al. (2021) and Zhong et al. (2020a) provide a current overview about machine learning in the law domain.

3. Legal Reasoning and Machine Learning

Most tasks in the legal domain – such as law matching – require some form of legal reasoning. In this section, we highlight some unsolved issues when using current machine learning models for such tasks.

Legal reasoning is the process of interpreting legal rules and applying them to facts (MacCormick, 1978). A legal rule is a *hypothetical imperative* (Engisch, 2005), which describes a conditional consequence. Given some facts and a law, legal reasoning is the process a human applies to understand whether the conditions of the law are met so that the consequence holds. This process requires a human to look up references and definitions, interpret terms, subsume facts under definitions, and conduct appreciation of conflicting values.

As of yet it is unclear if and to what extent statistical models can approximate this process. Bommasani et al. (2021) describe how the large language model GPT-3 fails with a simple task of inferring that a person is not entitled to a one million dollars compensation for a car when provided with the rule that damages are not enforceable if they are exorbitant. Holzenberger et al. (2020) conclude that a BERT model performs poorly in a inference task based on tax code rules. Zhong et al. (2020b) present a huge question answering data set from the National Judicial Examination in China, that requires advanced reasoning skills such as word matching, concept understanding, and multi-hop reading. They report only about 28% accuracy and low performance for many of these skills. These findings are in line with research about the limited capability of BERT models to reason (Rogers et al., 2020), e. g., about the physical world (Forbes et al., 2019), in reading comprehension (Jia and Liang, 2017), in logical tasks (Helwe et al., 2021), or over several inference steps (Richardson and Sabharwal, 2020; Zhou et al., 2020).

Many tasks in the legal domain that require such reasoning are modeled as entailment (Holzenberger et al., 2020; Zhong et al., 2020b; Rabelo et al., 2021). Entailment, also called “recognising textual entailment” or “natural language inference” (Poliak, 2020) is the task of inferring whether a text entails a hypothesis (Dagan et al., 2006). For our case, such a task could be: Given a law and a claim, does the law entail that the claim is true? We have identified several challenges with this approach. We will present them in the following subsections. We give examples which are based on real-world claims and laws in German. However, for easier comprehension we have translated and edited them.

3.1. References

Often legislative text contains references to other parts of the bill, or even to different bills. Consider Example 1:

Example 1

§ 9a SARS-CoV-2-Infektionsschutzverordnung (13.11.20)
(1) The isolation pursuant to § 8 (1) shall end at the earliest on the fifth day after entry if a person has a negative test result with regard to infection with the SARS-CoV-2 coronavirus.
Claim
Anyone entering from a risk area is in isolation for at least five days.

The claim is only true if the referenced isolation in § 8 (1) is imposed on people coming from an area of risk. § 8 (1) includes the following text:

Example 2

§ 8 SARS-CoV-2-Infektionsschutzverordnung (13.11.20)
(1) Persons who enter the Land of Berlin by land, sea or air from abroad and who have stayed in a risk area as defined in paragraph 4 at any time within 14 days prior to entry are obliged to go directly to their own home or other suitable accommodation immediately after entry and to seclude themselves there permanently for a period of 14 days after entry.

Since the article in Example 2 imposes the isolation for people from areas of risk, the claim is true. That means § 9a (1) entails the claim, *if* the content of § 8 (1) is known. If not, the hypothesis would be cast in doubt: § 8 (1) could order this specific kind of isolation for other reasons than return from an area of risk.

For the entailment task that means we have to provide both law sections and the claim as input during inference time. That is a problem, however, since a section can have several references, and the referenced sections can be far longer. Many Transformer-based models have a restricted input size of 512 tokens[3]. For example, the section § 4 SARS-CoV-2-EindmaßnV alone is 1122 token long in its encoded form[4]. So for any non-trivial reasoning task, it is impossible to encode all relevant articles as input.

A lawyer would resolve these references step-by-step. If we accept that it is in many cases impossible to encode all relevant information for one inference step, it would make sense to model the problem as a multi-step process. This has been done by Holzenberger and Van Durme (2021) who describe the process of resolving references and their logical connections, as well as mapping entities between the sections as *structure extraction*. Wolfson et al. (2020) break down (non-domain-specific) questions into atomic questions, which are then solved step-by-step.

3.2. Legal Terms and Definitions

Legal terms and definitions are similar to explicit references in legislation. Consider Example 3:

Example 3

§ 1 Zweite Krankenhaus-Covid-19-Verordnung (28.02.21)

This act shall apply to all hospitals licensed in the State of Berlin.

§ 3 Zweite Krankenhaus-Covid-19-Verordnung (28.02.21)

(1) Patients are allowed to receive once a day by one person for one hour visit.

Claim

Patients in psychiatric clinic are allowed to receive visitors only once a day.

Is the claim entailed by the laws? It depends on whether a psychiatric clinic is a licensed hospital. It is. A lawyer would find this out by either looking up a definition in another act, or in a commentary, or by interpreting the law (e. g., by arguing that other subsections in this act explicitly mention psychiatric clinics). However, it is not always necessary to do this process by yourself, since there are many commentary and journal articles by law professors and practitioners, as well as court decisions, that provide a definition and interpretation for such terms. In these cases legal terms are closely related to references in the sense that the definition of the term can be found somewhere else. So they can be understood as an *implicit reference*. However, their contents are more difficult to find than explicit references, since explicit references make it clear where and that the referenced information can be found.

Here, basically the same thoughts apply as with explicit references – a language model should have access to the respective information in order to make correct inferences.

3.3. Applicability

Applicability is a central concept in law: While some laws are universally applicable, most are limited in their scope, meaning that they only apply for certain conditions. Consider Example 4:

Example 4

§ 6 Zweite Krankenhaus-Covid-19-Verordnung (28.02.21)

(1) Accredited hospitals may perform scheduled admissions, provided that reservation and hold-free requirements are met and necessary staff resources and protective equipment are available.

Claim

Psychiatric clinics may only admit new patients if they comply with hold-free requirements and have sufficient staff and protective equipment.

Does the subsection entail the claim? Given that a psychiatry is a hospital, it does seem so. However, let's also consider § 6 (3):

Example 5

§ 6 Zweite Krankenhaus-Covid-19-Verordnung (28.02.21)

(3) Subsections 1 and 2 do not apply to psychiatric clinics.

§ 6 (3) explicitly states that § 6 (1) does not apply to psychiatric clinics. However, since this norm is not part of the text of the original example, a person would still infer from Example 4 that the claim is true. So if the goal is to infer the correct legal conclusion, *we must include all norms that are relevant for the applicability of a rule*. A limitation in scope could theoretically be anywhere: In a different section of the act, e. g., at the beginning, or even in a completely different act. It could also derive from a legislative competence, for example when the act is from a different state.

The entailment task only considers information that is contained in the text, and common knowledge. The scope of a legal norm is not common knowledge. Thus, if not explicitly in the text, all sections in the text are assumed to be in scope. That is a limitation of the entailment task. It does severely limit the correctness of a inference over the presented examples here.

3.4. Changing Rules and Amendments

Law changes. This does pose specific challenges for entailment. Consider Example 6:

Example 6

Claim

Also new: The ban on drinking alcohol outdoors now only applies in green areas as well as in parking lots. The ban for the public space as a whole is deleted.

Example 6 claims that a rule has been abolished. How can the veracity of such claim be determined? The relevant act here completely replaces the old act so one

[3] This restriction comes from the self-attention mechanism, whose computational and memory resources grow quadratically with sequence length (Vaswani et al., 2017). Beltagy et al. (2020) try to address this issue with the Longformer architecture.

[4] Encoded with the gbert-large tokenizer and in the version of the act from 22.03.20.

must check that 1) the rule was part of the former version of the act and 2) the rule is not present in the current act (also not in a different section or with a different wording). This requires access to the full text of both acts, and advanced reasoning skills.

3.5. Conclusion

We have seen that in order to do legal reasoning, access to a lot of information is needed. It is easy to find examples where encoding all relevant information in the input of a model is impossible due to restricted input size. And even if all information is present, a series of reasoning steps have to be conducted to arrive at the correct solution. As long as (BERT-based) language models are not able to handle negation (Helwe et al., 2021) or infer several reasoning steps (Zhou et al., 2020), there is little hope in expecting language models to do proper legal tasks. For the problem of changing rules (Section 3.4) it is even harder to conclude whether a claim such as in Example 6 is true.

We believe that these issues have to be addressed in order for language models to take on complex legal reasoning tasks. In that sense we agree with Holzenberger and Van Durme (2021) that the classical entailment task is an under-complex model for reasoning tasks about statutory applicability. Current research about multi-text modeling could be a basis for a more sophisticated model for such a task, as the highlights in our examples show how the same concepts are referenced over several subsections. For example, Ernst et al. (2021) align elements ("propositions") over summary/source documents.

4. Claim Extraction

In this section we present the claim extraction task, which labels claims about COVID-19 rules in German news articles.

4.1. Task Definition

The claim extraction task is concerned with detecting claims about legal rules in a document. A **claim** is "a statement that something is true or is a fact, although other people might not believe it".[5]

A claim as used in our data set is *any statement whose veracity can be determined with help of COVID-19-related legislation*. This means the statement must include a **legal consequence**, meaning *an imperative or a prohibition that is imposed by the law*. It excludes questions, suggestions, or plans. Its nature of being a claim comes from the fact that uttering a statement about a legal imperative does contain the implicit assertion that corresponding legislation exists, as in Example 7.

Example 7

Claim
Staying in the public space is allowed only alone or with another person or in the circle of the members of the own household.

Alongside the legal consequence a claim also contains **relevant context**, that is *any part of the document that includes a condition of the referred consequence*. However, the task considered in this paper has an important limitation: Only local context is part of the claim. We discuss this in the next section.

4.2. Claim Context

In order to make the task more tractable for this work, we made an important limitation: We only considered *relevant context that is located directly next to the legal consequence*. Thus, any claim is a continuous span of words. This is a significant limitation, since in many articles, the relevant context is distributed over the entire text.

For illustration, consider the case of an newspaper article where the introduction or the headline makes it clear that the following text only applies to restaurants. Then follows a list of legal rules, where the restaurant context is not explicitly stated again. The claim in Example 8 is about restaurants, but could also be about operas or theaters.

Example 8

Claim with missing context
"Employees and guests who are not at their seats must wear a medical face mask."

An optimal model would be able to extract the relevant context for a legal consequence from anywhere in the text. However, this is a genuine hard problem, since the question of whether a fact is actually a relevant condition requires knowledge of the law. The only way it can obtain this information is via inference from the training data, but this would mean it is impossible for the model to generalize well for more diverse problems. This is what makes this problem so difficult.

On the other hand, in cases where condition and consequence are located directly next to each other, this relevance can be determined fare more easily. In these cases there are often also syntactic hints that imply relevance ("If you visit the zoo, *then* you need to wear a mask"). We concentrate on local context in this work.

4.3. Technical Implementation

For the technical implementation, we conceptualize claim extraction as a sequence labeling task, more precisely as a token classification task using the BIO labelling scheme.

We use two model architectures for the claim extraction task: BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020). For both models we use the Huggingface transformers library (Wolf et al., 2020). We use model weights pre-trained on German corpora (Chan et al.,

[5] https://dictionary.cambridge.org/dictionary/english/claim. This definition must be distinguished from the legal concept of a claim ("a right to have something or get something from someone").

2020). Notably, the models are pre-trained on the Open Legal Data corpus (Ostendorff et al., 2020), which consists of German laws and court decision. Thus, we can consider the models as domain-specific.

4.4. Data Set

We annotate claims in newspaper articles and press releases. Thus, our claims and the articles are *natural samples*, that all occurred in the real world, and not *synthetic*. All annotations are done by one student with a law and a computer science background, so the choice of the articles might be biased towards personal preferences of the annotator.[6]

The claim extraction data set consists of articles with spans that represents claims. During pre-processing, the annotated articles are chunked into smaller pieces, due to the restricted input size of our models of 512 tokens. This results in our sample size being about 65% larger than the count of annotated articles.

We annotate 48 articles, which result in 79 samples. Every sample is a piece of text with spans that label claims. In total, the 79 samples contain 451 claims. Figure 1 presents the distribution of claims per sample. Approximately 59% of samples have a length of around 2500 chars, the remaining ones are shorter. The 48 articles are from 19 different websites. For example, 7 articles were from berlin.de and rbb24.de, and 6 articles from morgenpost.de. There 16 remaining websites account for 1-4 annotated articles each.

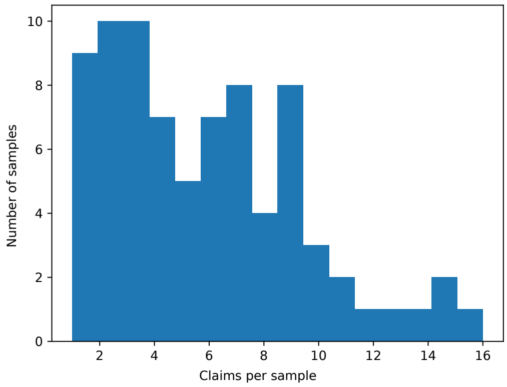

Figure 1: Claims per sample. Most samples contain less than 10 claims. Over 90% of the claims were shorter that 300 characters

4.5. Evaluation

We report claim extraction scores, conduct a manual analysis of the results, and discuss them.

Due to the small data set, we choose a 5-fold validation to account for variance between random training/test splits. We report the standard deviation between the five folds. All models were trained with batch size 30, learning rate 2^{-5}, and weight decay of 0.01.

Table 1 shows that gelectra-large yields with 0.467 the best F1 score for claim extraction.

4.6. Manual Analysis

We analyze the test set predictions of gbert-large and gelectra-large by hand. For that we evaluate both models by visualizing the results with color codes. The files can be found at our Github repository.

For analysis, both models obtain similar F1 score (0.468 for gelectra-large and 0.446 for gbert-large). In general, we find most true positives useful for the downstream law matching task, especially those from gelectra-large. Next, we highlight the most interesting results of this analysis.

4.6.1. Continuous Labeling of Spans

We observe that gelectra-large is better in labelling continuous spans of tokens as a claim, while gbert-large often inferred seemingly random different labels during a sentence, a pattern that is not present in the training data. This is the main contributor in why we perceive the results from gelectra-large as higher quality, despite similar F1 score.

After a sampling a subset of the test data, we estimate that gbert-large infers about 45% more often such "mixed label" sentences (13 vs. 29 occurrences). We do not know why this is the case. It could be that the different pre-training task for ELECTRA enables it to learn inter-label relationships more efficiently.

4.6.2. Claim Boundaries

Our articles often contain several claims next to each other. However, the ground truth and the model often disagree where a claim started, and where it ends.

Already during the annotation process we found it difficult to split up the sentences into different claims. Our guideline was: Make claims as small as possible, but make them bigger if this would result in lost context (see Section 4.2 for more details on the challenges of claim context).

We notice that the models have a tendency to infer shorter claims. About 60% more often do both models infer a new claim where the ground truth does not than the other way around. We notice that gelectra-large is substantially better than gbert-large in starting a sequences of I tokens with a B token, while gbert-large has a far more occurrences of OI tokens, a pattern not present in the ground truth.

4.6.3. False Positives

False positives are spans that the models infers as claims, but they are not labeled as a claim in the ground truth. We observe the following: Both models are particularly bad in inferring that plans of politicians to pass new laws are not claims. A large amount of the total false positives fall into this category. Example: "Relaxations of the rules are also planned in the sports sector."

[6] Annotation guideline: https://raw.githubusercontent.com/DFKI-NLP/covid19-law-matching/master/annotation_guideline.pdf

Model	F1	Precision	Recall
gbert-base	0.398 (± 0.051)	0.333 (± 0.048)	0.496 (± 0.069)
gbert-large	0.429 (± 0.057)	0.354 (± 0.052)	**0.547** (± 0.067)
gelectra-base	0.418 (± 0.036)	0.357 (± 0.034)	0.507 (± 0.059)
gelectra-large	**0.467** (± 0.064)	**0.417** (± 0.073)	0.535 (± 0.052)

Table 1: Results for the claim extraction task measured as F1 score, precision, and recall including standard deviation over five folds. gelectra-large yields overall the best results.

While such a statement constitutes a claim in the classical sense (as defined by the Cambridge Dictionary), for this work we only consider such claims whose veracity can be determined with help of COVID-19-related legislation (see Section 4.1), which is not the case here. We also notice that the models often correctly infer claims that are erroneous not labeled as such in the ground truth. This is the case in 33% of all false positives for both models (34 inferred claims for gelectra-large, 32 for gbert-large). Since the data set contains a total of 451 claims, these missing claims constitute at least 7% of all claims.

4.6.4. False Negatives

False negatives (model does not infer a claim that is labeled as such in the ground truth) are far more rare than false positives. About 50% of all samples do not contain a false negative (40 for gelectra-large and 37 for gbert-large), while the rest does contain mostly around 1-2 instances. We also find that the annotation quality is much higher in this regard, since we do not identify a false negative that should not have been labeled as a claim in the ground truth.

4.6.5. Headings

In the original articles, often a heading precedes a claim. The models are inconsistent in including them in a claim. However, this inconsistency is also present in the ground truth. Our annotation guideline does not contain clear rules whether to include the headings or not in a claims. Another issue is that when text is converted from HTML to plaintext, the semantic information of headings is lost: The heading becomes just a piece of text, often a grammatically incorrect sentence without punctuation.

4.7. Discussion

We find it conceptually and technically challenging to extract the full context of a claim (Section 4.2). In addition, when many claims are contained in one document, it is difficult to exactly determine the boundaries of a claim. This is also reflected in our results, where there is a high disagreement between models and ground truth on claim boundaries (Section 4.6). One potential solution for this issue could be to only annotate the legal consequence, and supply the full article as context. Then the model could learn out what is relevant, and claim boundaries would be less challenging. However, this is also not a silver bullet. First, conceptually it is often difficult to determine what is part of the legal consequence, and what is part of the conditions (Engisch, 2005). Second, providing the full article next to the legal consequence is a problem where the input size of the model is restricted, as it is with the state-of-the-art Transformer architectures. Third, since the articles mostly contain multiple claims (see Figure 1), and the relevant conditions are not annotated, the full article as context will be noisy. Hence, the data set probably would need to be far bigger to account for that.

We find that gelectra-large is the most suited for the task. Even when compared to a gbert-large with very similar F1 score, it performed better in some important aspects. gelectra-large labels continuous spans better and respects claim boundaries better (i. e., starting a claim with a B token). In one of our experimental setups, gelectra-large shows substantial better F1 score than the other models.

While we found that some claims in the ground truth were not labeled as such, this affected only about 7% of all claims (Section 4.6). In addition, the labeled claims in the ground truth appear to be of high quality: During review we did not found any labeled claim that was misclassified as such.

In general, manual analysis of the results are promising. We believe the results indicate that a model with sufficient good real-life performance can be devised.

5. Law Matching

In this section, we present the law matching task, which infers whether a given subsection[7] applies to a claim. The goal of the law matching task is to match any subsection of a law that helps a human to determine whether the claim is correct or not.

5.1. Task Definition

The law matching task is concerned with matching a subsection of a law with a claim. **Given a set of subsections S that apply to a claim c, any subsection $s \in S$ matches c.** Importantly, opposed to entailment, we do not require the subsection to entail the claim

[7] We match on subsections. In principal, this also works with whole sections or even single sentences. However, while single sentences would lack important context, whole sections are often too long for the restricted input size of the models we use.

on its own (see Section 3 for the underlying problem). However, the entire set of matching subsections do entail the claim. To illustrate, the Example 9 constitutes a valid match.

Example 9

Claim
Sport in covered sports facilities is allowed if it is essential. This includes equestrian sports within the scope of animal welfare considerations.
Subsection
The practice of sports in covered sports facilities, fitness and dance studios and similar facilities is only permitted insofar as it is necessary
1. for the sport of the group of persons mentioned in paragraph 1,
2. for equestrian sports to the extent that is absolutely necessary from the point of view of animal welfare,
3. for therapeutic treatments as well as uses in accordance with paragraph 1.
Otherwise, it is prohibited.

5.2. Scope

We define the scope of the law matching task in order to make it more tractable:

- We leave out claims about changing rules (see Section 3.4). This is a serious limitation, since those claims became more prevalent when rules are eased during the second part of 2021.
- We leave out claims that can only be verified with the fact of the nonexistence of a law. The problematic here is similar to the one with changing laws.
- We consider only laws from the Berlin state (Germany), and claims about laws from Berlin.
- We leave out claims which are missing relevant context (see Section 4.2).

5.3. Technical Implementation

We use the same models as in the claim extraction task (Section 4.3), except that the models use a sequence classification head.

As baseline we use a classification based on cosine similarity between TF-IDF vectors. Since matching subsections and claims often use similar words, it can expected that they have a higher degree in cosine similarity than non-matching ones. Both claims and subsections are transformed to TF-IDF vectors using the Scikit-learn library[8]. Then the cosine similarity for every claim-subsection pair is calculated, which, together with the corresponding label, was used as data.

5.4. Data Set

For the law matching task we provide a data set with a total of 858 samples. To create that, we annotate 328 claims with subsection that together justify the claim. The claim paired with every subsection separately become a positive sample, and for every positive sample a random subsection is chosen as a negative sample. Thus, the data set is balanced with 50% positive and 50% negative samples. The claims are from March 2020 to July 2021. They are not equally distributed over this time, as 57% of the claims reference a time between March to June 2021.

The text for the subsections originates from an additional data set for the COVID-19-related legislation in Berlin. We collected 13 COVID-19-related acts from the official website http://gesetze.berlin.de, with a total of 975 sections[9], for which the applicability period is provided.

5.5. Evaluation

We report scores for this task (Table 2) and discuss them (Section 5.6). Due to the small data set, we choose a 5-fold validation to account for variance between random training/test splits. In parentheses we report the standard deviation over the five folds.

All models are trained with batch size $b = 3$, learning rate $\lambda = 1^{-5}$, and weight decay of $d = 0.01$.

As shown in Table 2, gelectra-large outperforms all other models. gbert-large has a lower F1 than gbert-base, and a far higher standard deviation.

5.6. Discussion

The large German BERT model gbert-large shows a worse performance than the other Transformer-based models, on par with the TF-IDF baseline. Our analysis shows that gbert-large does not learn the 50%-50% distribution of the original data set (it only infers a match in 38% of the samples). However, this could be due to a random error. For small data sets, the randomly initialized weight of the classification head can result in high variance in classification performance (Risch and Krestel, 2020).

Manual analysis of the results of the baseline model shows that TF-IDF performs poorly in matching claims with subsections that contained long lists (i. e., having lots of words), and with subsections that contained vastly different wording. This can be explained by TF-IDF vectors being cosine similar if they contain similar words which are similar frequent over the corpus of documents. Since this is not necessarily the case, Transformer-based models are more suited for this task, which is also shown through their better results.

In general, the results are promising, with gelectra-large achieving up to 0.91 F1 score. However, the data set is relatively small, and skewed to a period between March and June 2021 (Section 4.4), and we limit the scope of our task (Section 5.2). Thus, the results do not indicate how well the models generalize to different time spans, or similar tasks.

[8] https://scikit-learn.org/stable/

[9] A section may consist of several subsections.

Model	F1	Precision	Recall
TF-IDF baseline	0.741 (± 0.094)	**0.944** (± 0.016)	0.619 (± 0.139)
gbert-base	0.880 (± 0.060)	0.871 (± 0.091)	0.893 (± 0.047)
gbert-large	0.736 (± 0.417)	0.723 (± 0.420)	0.757 (± 0.424)
gelectra-base	0.857 (± 0.047)	0.883 (± 0.087)	0.836 (± 0.034)
gelectra-large	**0.914** (± 0.070)	0.895 (± 0.120)	**0.941** (± 0.024)

Table 2: Scores for the law matching task measured as F1 score, precision, and recall including standard deviation over five folds. gelectra-large yields overall the best results.

We have discussed, in Section 3, the difficulties of legal reasoning. The fact that our models perform well anyway can be explained with the scope of the task: The COVID-19-related legislation does contain a lot of concrete, self-contained rules and restrictions, in differentiation to more abstract and modular rules. This is visible in the fact that around 75% of claims are annotated with a single subsection, which means that the referenced rule is contained in one single place. Also the simple baseline achieves a F1 score of 0.741, which indicates that simple word similarity is enough to achieve a good performance on this dataset. Also, we limit the scope as explained in Section 5.2, and the negative samples are constructed with random subsections. Both make the task easier.

It is also important to note that our task definition (Section 5.1) is different from the classic entailment task, in which many tasks are mostly modeled. Notably, in the entailment all relevant information (except "common knowledge") must be contained in the input. In contrast, our task allows for matching subsections that do apply, but where the applicability criteria is missing in the sample (e. g., because they are located in a different section of the law). We show that – at least for the restricted domain of COVID-19-related legislation in Berlin – this can work. However, it is questionable whether models can generalize well in such a setup for a more diverse and broader task.

6. Conclusion

Legal reasoning is difficult to model for current NLP architectures. We show by example (Section 3) that several crucial steps in the reasoning process are underspecified (e. g., because of restricted input size and access to information) or that models show generally bad performance in those (e. g., multi-hop reasoning, negation) and that entailment is not a sufficient model in the reasoning task. Thus we conclude that inference over legal conclusions is only possible in simple cases with a limited scope.

For claim extraction (Section 4), we find that gelectra-large is suited best for the task, with a F1 score of 0.467. Even when compared to gbert-large with similar F1 performance, gelectra-large learns better to extract sentences as a whole, and to start claims with a B token according to our labeling scheme. We find the manually inspected results promising, especially given the small data set size.

However, we find it difficult to extract the full relevant context from the article. Thus, we only extract "local" context (Section 4.2), which result in some claims with missing context. Solving this is a conceptual and technical challenge. In addition, the data set is currently relative small and about a specific period during the pandemic. Extending the data set will be subject of future work.

For law matching, gelectra-large obtains a F1 score of 0.91, indicating that for this particular domain, matching the relevant laws is possible with sufficient accuracy. This comes with the limitation that there are several classes of claims for which it is very difficult to obtain evidence, and we have excluded those from the law matching task. These include claims about changes in legislation (Section 3.4) and claims where the evidence is the absence of a law. Since such claims are part of our claim extraction data set, this means the performance of claim extraction and law matching task will affect each other. Moreover, the fact that the negative samples are constructed randomly also makes the task less challenging, which is shown by the simple, but strong baseline.

The experimental results we obtained are promising and we believe our work can serve as a first step towards a full claim checking pipeline for COVID-19-related legislation, or similar applications. Especially the data set that we publish along with this paper can be the basis for future research. For that, a valuable future contribution would be solutions for the difficult cases of claims which we have excluded in this work (see Section 5.2). In addition, we believe the challenges outlined in Section 3 are a limitation for more solving more complex legal tasks with machine learning. A more complex model than entailment for legal tasks would hence be a valuable contribution.

7. Acknowledgements

The research presented in this article is partially funded by the German Federal Ministry of Education and Research (BMBF) through the projects QURATOR (Rehm et al. (2020); Unternehmen Region, Wachstumskern, no. 03WKDA1A) and PANQURA (no. 03COV03E).

8. Bibliographical References

Atanasova, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Martino, G. D. S., and Nakov, P. (2018). Overview of the CLEF-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *CoRR*, abs/1808.05542.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv:2004.05150*.

Beltran, J. and Larraz, I. (2021). Claimhunter: An unattended tool for automated claim detection on twitter. In Konstantin Todorov, et al., editors, *Proceedings of the 1st International Workshop on Knowledge Graphs for Online Discourse Analysis*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. (2021). On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Chakrabarty, T., Hidey, C., and McKeown, K. (2019). IMHO fine-tuning improves claim detection. 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 558–563, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November. Association for Computational Linguistics.

Chalkidis, I., Jana, A., Hartung, D., II, M. J. B., Androutsopoulos, I., Katz, D. M., and Aletras, N. (2021). Lexglue: A benchmark dataset for legal language understanding in english. *CoRR*, abs/2110.00976.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, et al., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Engisch, K. (2005). *Einführung in das juristische Denken*. W. Kohlhammer, 10st edition.

Ernst, O., Shapira, O., Pasunuru, R., Lepioshkin, M., Goldberger, J., Bansal, M., and Dagan, I. (2021). Summary-Source Proposition-level Alignment: Task, Datasets and Supervised Baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California, June. Association for Computational Linguistics.

Forbes, M., Holtzman, A., and Choi, Y. (2019). Do neural language representations learn physical commonsense? *CoRR*, abs/1908.02899.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., and Derczynski, L. (2019). SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Hanselowski, A., Stab, C., Schulz, C., Li, Z., and Gurevych, I. (2019). A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China, November. Association for Computational Linguistics.

Helwe, C., Clavel, C., and Suchanek, F. M. (2021). Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

Holzenberger, N. and Van Durme, B. (2021). Factoring Statutory Reasoning as Language Understanding Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2742–2758, Online. Association for Computational Linguistics.

Holzenberger, N., Blair-Stanek, A., and Durme, B. V. (2020). A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, et al., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

MacCormick, N. (1978). *Legal Reasoning and Legal Theory*. Oxford University Press, USA.

Marín, M. J. (2017). Legalese as Seen Through the Lens of Corpus Linguistics. An Introduction to Software Tools for Terminological Analysis. *International Journal of Language & Law (JLL)*, 6(0), August. Number: 0.

Ostendorff, M., Blume, T., and Ostendorff, S. (2020). Towards an open platform for legal information. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, Aug.

Ostendorff, M., Ash, E., Ruas, T., Gipp, B., Moreno-Schneider, J., and Rehm, G. (2021). Evaluating document representations for content-based legal literature recommendations. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, page 109–118, New York, NY, USA. Association for Computing Machinery.

Pankovska, E., Schulz, K., and Rehm, G. (2022). Suspicious Sentence Detection and Claim Verification in the COVID-19 Domain. In *Proceedings of the 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022)*, 04.

Poliak, A. (2020). A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online, November. Association for Computational Linguistics.

Rabelo, J., Goebel, R., Yoshioka, M., Kano, Y., Kim, M.-Y., Yoshioka, M., and Satoh, K. (2021). Summary of the competition on legal information extraction/entailment (coilee) 2021. In *COLIEE'21*, Online, June.

Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J. M., Ostendorff, M., Zaczynska, K., Berger, A., Grill, S., Räuchle, S., Rauenbusch, J., Rutenburg, L., Schmidt, A., Wild, M., Hoffmann, H., Fink, J., Schulz, S., Seva, J., Quantz, J., Böttger, J., Matthey, J., Fricke, R., Thomsen, J., Paschke, A., Qundus, J. A., Hoppe, T., Karam, N., Weichhardt, F., Fillies, C., Neudecker, C., Gerber, M., Labusch, K., Rezanezhad, V., Schaefer, R., Zellhöfer, D., Siewert, D., Bunk, P., Pintscher, L., Aleynikova, E., and Heine, F. (2020). QURATOR: Innovative Technologies for Content and Data Curation. In Adrian Paschke, et al., editors, *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*, Berlin, Germany, 02. CEUR Workshop Proceedings, Volume 2535. 20/21 January 2020.

Richardson, K. and Sabharwal, A. (2020). What does my QA model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

Risch, J. and Krestel, R. (2020). Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France, May. European Language Resources Association (ELRA).

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs]*, November. arXiv: 2002.12327.

Schneider, J. M., Rehm, G., Montiel-Ponsoda, E., Rodríguez-Doncel, V., Martín-Chozas, P., Navas-Loro, M., Kaltenböck, M., Revenko, A., Karampatakis, S., Sageder, C., Gracia, J., Maganza, F., Kernerman, I., Lonke, D., Lagzdins, A., Gil, J. B., Verhoeven, P., Diaz, E. G., and Ballesteros, P. B. (2021). Lynx: A Knowledge-based AI Service Platform for Content Processing, Enrichment and Analysis for the Legal Domain. *Information Systems*, page 101966. Special Issue on Managing, Mining and Learning in the Legal Data Domain.

Siewert, M., Wurster, S., Messerschmidt, L., Cheng, C., and Buthe, T. (2020). A German Miracle? Crisis Management During the COVID-19 Pandemic in a Multi-Level System. SSRN Scholarly Paper ID 3637013, Social Science Research Network, Rochester, NY, June.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*, December. arXiv: 1706.03762.

Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June. Association for Computational Linguistics.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. *arXiv:2004.14974 [cs]*, October. arXiv: 2004.14974.

Wang, W. Y. (2017). ”Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wolfson, T., Geva, M., Gupta, A., Gardner, M., Goldberg, Y., Deutch, D., and Berant, J. (2020). Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020a). How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020b). Jec-qa: A legal-domain question answering dataset. In *Proceedings of AAAI*.

Zhou, X., Zhang, Y., Cui, L., and Huang, D. (2020). Evaluating commonsense in pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9733–9740, 04.