Artificial Intelligence and Law (2024) 32:839–862
https://doi.org/10.1007/s10506-023-09369-4

ORIGINAL RESEARCH

I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets

Daniel Braun[1]

Accepted: 12 June 2023 / Published online: 27 June 2023

Abstract
Legal documents, like contracts or laws, are subject to interpretation. Different people can have different interpretations of the very same document. Large parts of judicial branches all over the world are concerned with settling disagreements that arise, in part, from these different interpretations. In this context, it only seems natural that during the annotation of legal machine learning data sets, disagreement, how to report it, and how to handle it should play an important role. This article presents an analysis of the current state-of-the-art in the annotation of legal machine learning data sets. The results of the analysis show that all of the analysed data sets remove all traces of disagreement, instead of trying to utilise the information that might be contained in conflicting annotations. Additionally, the publications introducing the data sets often do provide little information about the process that derives the "gold standard" from the initial annotations, often making it difficult to judge the reliability of the annotation process. Based on the state-of-the-art, the article provides easily implementable suggestions on how to improve the handling and reporting of disagreement in the annotation of legal machine learning data sets.

Keywords Data annotation · Legal corpora · Annotator agreement

1 Introduction

Disagreement is the default state in legal proceedings. While it is often their goal to settle a disagreement, e.g. by a court decision, the state of disagreement can prevail for a long time and in some cases, a disagreement might never be settled. Parties in a lawsuit can disagree, legal scholars can disagree, courts can disagree with each other, and even judges in the same court can disagree. Sometimes, the state of disagreement is so valuable to one of the involved parties, that they are willing to pay

✉ Daniel Braun
d.braun@utwente.nl

[1] Department of High-Tech Business and Entrepreneurship, University of Twente, Hallenweg 17, 7522 NH Enschede, The Netherlands

large sums in out-of-court settlements, to prevent the official resolution of the underlying disagreement.

The process and methods surrounding artificial intelligence (AI) and machine learning (ML), on the other hand, are optimised towards finding a single "truth", a gold standard. From the annotation of data sets, where "outliers" are often eliminated by majority vote and a high inter-annotator agreement is seen as a sign of quality, to the presentation of the predictions models make, where often, only the most probable output will be shown, disagreement is systematically eradicated in favour of a single "truth".

When legal machine learning data sets are annotated, these opposing worlds have to be combined. Annotations have to be correct from a legal perspective and usable from a technical perspective. This article presents an analysis of how disagreement between annotators is handled in the annotation of legal data sets. A review of 29 manually annotated machine learning data sets and the corresponding publications describing their annotation shows that most data sets are annotated by multiple annotators and that the majority of the accompanying papers describe how disagreement is handled, however, often only in little detail. None of the data sets we investigated provides the raw data including disagreeing annotations in the published corpus.

In their paper "Analyzing Disagreements", Beigman Klebanov et al. (2008) differentiate two types of disagreement in annotations: mistakes due to a lack of attention and "genuine subjectivity". Disagreement that originates from a lack of attention does arguably not provide any valuable information and should therefore be removed from corpora. Identifying whether a disagreement originates from a lack of attention or not is relatively easy when involving the original annotator.

Genuine subjectivity in a corpus, on the other hand, can be valuable. It can help to provide a more balanced picture of the subject matter. In the annotation of legal data sets, e.g., the annotation of contract clauses, subjectivity can, for example, arise from different interpretations of vague legal terms (Li 2017). If there is no legal precedent for what a "sufficient timespan" is in a given context, annotators can have different, subjective, interpretations. And even if there is legal precedent, it is up for the interpretation, and therefore subjectivity, of the annotators to assess, whether the contexts are similar enough for the precedent to be applicable.

Beigman Klebanov et al. (2008) analyse disagreement in a corpus for metaphor detection. While in this case, all disagreement can be explained with either a lack of attention or subjectivity, we believe that, in the legal domain, there can also be disagreement that is neither based on a mistake, nor on subjectivity. A more objective disagreement. Such disagreement can, for example, originate from missing information. When corpora are annotated, the data is often taken out of context, to a certain degree. When annotating whether a given contract clause is void, two annotators might draw different conclusions and both of them can be correct, depending on whether both parties of the contract are businesses or one party is a consumer. Another source for a more "objective disagreement" can be that different courts have made conflicting decisions which are both valid at the same time and annotators base their disagreeing annotations on different of these conflicting decisions. A recent example of such a situation can be found in Germany. In January

2022, the time for which a person is legally considered as "genesen" (recovered) from COVID-19, and therefore considered less likely to be infectious and exempted from some of the restrictions in place at the time, was shortened to 90 days. Subsequently, several urgent motions have been filed with administrative courts all over the country. While most of the courts, including the Verwaltungsgericht Berlin[1] and Verwaltungsgericht Hamburg,[2] ruled the law that enabled the change to be most likely unconstitutional, the Verwaltungsgericht Koblenz[3] decided otherwise. Like subjectivity, such more "objective" disagreement among annotators can also be very valuable and should therefore be represented in corpora.

After an analysis of the current practices surrounding disagreement during the annotation of legal data sets, Sect. 6 provides suggestions on how the handle different types of disagreement between annotators in the annotation process.

Legal data sets and tasks are very diverse and so are the possible reasons for disagreeing annotations. How exactly disagreement can be leveraged depends on these and other factors, like the expertise and diversity of annotators. However, the pure existence and extent of disagreement are already valuable information, as also pointed out by Prabhakaran et al. (2021). On an aggregated level, the extent of disagreement is already used as a measure for the quality of annotations, through metrics like inter-annotator agreement (see Sect. 5.3). Having information about disagreement available within a data set on an item level, instead of just an aggregated metric, can also be leveraged to provide alternative opinions, e.g. by presenting more than just one probable prediction or training multiple models based on individual annotators, and assessing the confidence with which a prediction can be made by a model trained on the data set. If it is ensured that the disagreement in the data set is based on genuine disagreement, rather than mistakes or a lack of attention, the disagreement can also provide valuable insights into whether a certain assessment might be especially difficult to make. Therefore, we believe that having genuinely disagreeing annotations represented in a data set is always valuable, at least for reasons of quality control and transparency.

On a task and data set specific level, in order to leverage the full potential of having access to disagreeing annotations, we believe that it is also necessary to provide information about the annotation process and the expertise of the annotators. Therefore, our suggestions in Sect. 6 do not just include advice on how to handle disagreement in the annotation process, but also how to report on the annotation process in more detail. If in addition to the disagreeing annotation, there is also information available about the annotators, like their expertise or the jurisdiction in which they practice, we can use this information to give different priorities to individual annotations based on the input that is used to make a prediction.

[1] VG Berlin, Decision 16.2.2022, 14L 24/22.

[2] VG Hamburg, Decision 14.2.2022, 14 E 414/22.

[3] VG Koblenz, Decision 23.2.2022, 3L 150/22 and 3L 169/22.

2 Related work

In their paper "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing", Basile et al. (2021) provide suggestions on how "to embrace a perspectivist stance in ground truthing", including involving a sufficient number of annotators, involve a heterogeneous group of annotators, and be mindful of the shortcomings of a majority vote decision. Basile et al. (2021) also specifically point out that including different perspectives can not just be valuable for tasks that are traditionally seen as subjective, but also for tasks that are traditionally seen as more objective, like medical decision-making. We believe that also applies to legal decision-making.

While considering how to handle different perspectives during annotation tasks that are widely considered "objective" is not yet widespread, for subjective annotation tasks, different approaches have been suggested by Rottger et al. (2022), Davani et al. (2022), Ovesdotter Alm (2011), and others. One domain in which particular emphasis has been put on the perspectives of and disagreement between annotators in recent years is the annotation of hate speech (Sachdeva et al. 2022; Kralj Novak et al. 2022; Akhtar et al. 2020). Arguably, the annotation of hate speech is a more subjective task. Most of the works in this area consider situations where there is a large number of annotators that is not necessarily highly qualified. In such situations, removing "noise", i.e. objectively wrong annotations, possibly caused by a lack of attention, has a high priority. Our analysis in Sect. 5.2 shows that legal data sets are often annotated by a small but highly skilled number of annotators. Although skilled annotators are not immune to mistakes caused by a lack of attention, it is reasonable to assume that disagreement between skilled annotators is more often actual disagreement than noise, in comparison with crowd-sourced annotations.

Literature that focuses on more objective tasks, e.g. in the medical domain, agrees that majority voting is not a favourable approach for handling disagreement during the annotation (Campagner et al. 2021; Sudre et al. 2019). However, only Sudre et al. (2019) suggest using labels from individual annotators, preserving the disagreement, while Campagner et al. (2021) suggest new ways of consolidating disagreeing labels to a single ground truth. Moreover, Sudre et al. (2019) do suggest just using raw individual annotations, without accounting for pseudo-disagreement, caused by a lack of attention or simple mistake during the annotation. We believe that it is important to remove this kind of noise to come to create a data set in which the presented disagreement provides valuable information.

On a more general level, in addition to Basile et al. (2021), Prabhakaran et al. (2021) and Jamison and Gurevych (2015) have investigated the question whether including disagreeing annotations in corpora is valuable or only adds noise. While Prabhakaran et al. (2021) conclude that "dataset developers should consider including annotator-level labels", without giving concrete advice on how to include disagreement in the annotation process, Jamison and Gurevych (2015) argue that "the best crowdsource label training strategy is to remove low item agreement instances from the training set". The latter is probably influenced by the fact that the authors specifically consider crowdsourcing, where, as mentioned before, disagreement can be more likely to be caused by a lack of attention than actual disagreement.

This article investigates how disagreement is handled in the annotation of legal data sets. Data sets that contain disagreeing labels can often not directly be used to train predictive models, because the standard training approaches for such models rely on a single ground truth. While not the focus of this work, technical approaches on how to utilise disagreeing labels in the training of predictive models already exist in the realm of hate speech detection, e.g. by using the individual annotations to train ensemble models which outperform a single model trained on the consensus annotation (Akhtar et al. 2020) or using the information about the uncertainty that can be derived from the disagreement in the annotations (Klemen and Robnik-Šikonja 2022; Ramponi and Leonardelli 2022).

3 Scope

As digitisation progresses within court systems and the legal domain at large, the number of available legal data sets is constantly increasing. However, a large share of these data sets is not annotated and therefore not relevant in the context of this work. Examples of such data sets include legislation from different countries, like the data provided by Open Legal Data (Ostendorff et al. 2020). Such unlabelled data can, for example, be used to train large language models in an unsupervised fashion (see e.g. Chan et al. (2020)). Parallel corpora containing the same (unlabelled) documents in multiple languages can also be used to train models for machine translation (Steinberger et al. 2006).

Another large share of existing legal data sets contains labels, however, these labels are not the result of a deliberate and manual annotation process, but an inherent part of the underlying data. Court decisions, for example, always contain the decision of the court. This information might not be available in a structured format and extracting that information can be a difficult task, yet, when building such a data set, annotators do not have to make a legal assessment. Therefore, such data sets are also not the focus of this work and we will consider them, for the purpose of this paper, as not (manually) annotated. Although out of the scope of this work, it is interesting to note that court decisions made by multiple judges, e.g. at the Supreme Court of the United States (SCOTUS) and the Federal Constitutional Court in Germany (Bundesverfassungsgericht, BVerfG), face a similar problem of how to handle disagreement, in this case between judges. Just like the data sets described in Sect. 4, they chose different approaches. At the SCOTUS, decisions are usually made individually by each judge and the outcome is then based on a majority vote, resulting in frequent dissenting opinions. The BVerfG, on the other hand, is focused on trying to achieve consensus through common deliberations of all judges before a vote is caste (Lübbe-Wolff 2022). In case of a unanimous decision, courts like the providers of data sets have to decide how to handle the dissenting opinion, e.g. whether individual votes are disclosed.

In this article, we investigate how disagreement is handled in the annotation of legal data sets within the scientific literature. Therefore, we will only consider data sets which have been created in a formal, deliberate annotation process, excluding data sets that have not been annotated or consist of only inherent labels, that can be directly extracted from the data itself. We also only consider annotations that include some kind of legal knowledge, excluding data sets that are, for example, labelled with only linguistic information, like part of speech, named entity recognition, and speech-to-text, because we want to focus on investigating the handling of disagreement in legal matters.

4 Data sets

Although the number of openly available legal data sets is constantly growing, very few of them are listed in traditional databases for language resources. The catalogue of the European Language Resources Association (ELRA),[4] for example, lists only 30 resources within the domain “law” (as of September 2022). All of these 30 resources are either unlabelled legislative texts or parallel corpora of legal texts in multiple languages, neither of which fits the scope of this article.

Therefore, we decided to use less traditional sources, so-called “awesome lists” on GitHub. An “awesome list” is a GitHub repository that consists of a curated list of resources for a specific purpose (Wu et al. 2017). Prominent exmaples for such lists include awesome NLP,[5] awesome Java,[6] and awesome machine learning.[7] These curated lists are also frequently used in scientific literature, for the three aforementioned see e.g. Sas and Capiluppi (2022), Gonzalez et al. (2020), and Zahidi et al. (2019).

We identified four such curated lists that are dedicated (at least partially) to legal data sets[8]:

- **awesome-legal-data:** The repository is curated by Schwarzer (2022) of the German NGO Open Justice e.V. and contains 24 data sets from 13 countries and the European Union.
- **Legal Text Analytics:** This list contains 56 data sets from 10 countries and the European Union and is maintained by Waltl (2022) of the German NGO Liquid Legal Institute e.V.
- **Must-read Papers on Legal Intelligence:** While this list is focused on the curation of papers, it also contains a curated list of 16 data sets in 8 languages, maintained by Xiao et al. (2021).
- **Datasets for Machine Learning in Law:** This repository contains 24 data-sets and is curated by Guha (2021).

Together, these lists contain 120 data sets, however, there are overlaps between the lists and, more importantly, only a small fraction of these 120 data sets fall within

[4] http://catalogue.elra.info.

[5] https://github.com/keon/awesome-nlp.

[6] https://github.com/akullpp/awesome-java.

[7] https://github.com/josephmisiti/awesome-machine-learning.

[8] All last accessed 01/12/2022.

Table 1 Number of different types of data sets in the repositories

Repository	Non-legal	Legal corpora with			Total
		No annot.	Non-legal annot.	Legal annot.	
Awesome-legal-data	0	22	1	1	24
Legal Text Analytics	3	40	3	10	56
Must-read Papers on Legal Int.	0	7	3	6	16
Datasets for ML in Law	0	14	4	6	24
Total	3	83	11	23	120

the scope described in Sect. 3. Table 1 shows an overview of the number of (legal) data sets in each repository and whether they contain annotations, and if so, which kind of annotations. The overview shows that out of the 120 entries, only 23 data sets fall within the scope of our analysis, i.e. are legal data sets that have been annotated with legal information. It is important to remember that for this article, we take a process perspective on annotation and only consider data sets as annotated that went through a manual, scientific annotation process in which a legal assessment was made by the annotators. Therefore, data sets without annotation in this case can also include data sets that contain labels and can be used for supervised machine learning, if these labels have not been generated in a manual annotation process. This includes data sets for legal outcome prediction that have the verdict as label or summarisation data sets that take a certain part of a document as a given summary.

Additionally, the repositories are not mutually exclusive, i.e. a corpus can be listed in multiple repositories. Therefore, the total number of distinct data sets in the four repositories, that fit the scope of our analysis, is 21. In order to widen our analysis, we also included data sets that were not directly listed in one of the four repositories but were cited by one of the papers listed in them. The LexGLUE data set by Chalkidis et al. (2022), for example, is listed in multiple repositories and does not introduce any newly annotated data sets that are within the scope of our work, but bundles six data sets that are partially relevant to this analysis and are themselves not listed in any of the repositories. In this way, eight additional data sets were added, contributing to a total of 29 data sets analysed. Table 2 lists the data sets, their corresponding publications, the language of the documents in the data sets, as well as the source from which the data set was gathered (ALD = awesome-legal-data, LTA = Legal Text Analytics, PLI = Must-read Papers on Legal Intelligence, MLL = Datasets for Machine Learning in Law, CIT = Citations).

The majority of data sets consist of English texts only (20 out of 29), followed by German and Standard Chinese (three each). Only two out of the 29 data sets contain multilingual data and only one contains french texts. While many unlabelled corpora in the legal domain are multilingual, because they are based on multilingual data provided by the European Union, annotated data sets rarely are. Most likely due to the high costs of legal annotations in general. Of the two multilingual corpora only one was really annotated in multiple languages. The corpus from Drawzeski et al.

Table 2 List of data sets analysed in this article

ID	Title	Author(s)	Lang	Source
1	A Case Study on Legal Case Annotation	Wyner et al. (2013)	EN	CIT
2	A Corpus for Multilingual Analysis of Online Terms of Service	Drawzeski et al. (2021)	DE, EN, IT, PO	CIT
3	A Dataset and an Examination of Identi-fying Passages for Due Diligence	Roegiest et al. (2018)	EN	LTA
4	A Statutory Article Retrieval Dataset in French	Louis and Spanakis (2022)	FR	LTA
5	A test collection for evaluating legal case law search	Locke and Zuccon (2018)	EN	PLI
6	Automating the Classification of Finding Sentences for Linguistic Polarity	Walker et al. (2020)	EN	LTA
7	BUILDNyAI	Tiwari et al. (2022)	EN	LTA
8	Cail2019-scm: A dataset of similar case matching in legal domain	Xiao et al. (2019)	CN	PLI
9	Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension	Duan et al. (2019)	CN	PLI
10	Classifying semantic types of legal sentences: Portability of machine learning models	Glaser et al. (2018)	DE	MLL
11	CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service	Lippi et al. (2019)	EN	MLL
12	Clause Topic Classification in German and English Standard Form Contracts	Braun and Matthes (2022)	DE,EN	CIT
13	Contract Discovery: Dataset and a Few-Shot Semantic Retrieval Challenge with Competitive Baselines	Borchmann et al. (2020)	EN	CIT
14	Corpus for automatic structuring of legal documents	Kalamkar et al. (2022)	EN	LTA
15	Cuad: An expert-annotated nlp dataset for legal contract review	Hendrycks et al. (2021)	EN	LTA
16	Design and Implementation of German Legal Decision Corpora	Urchs et al. (2021)	DE	LTA
17	ECHR: Legal Corpus for Argument Mining	Poudyal et al. (2020)	EN	LTA
18	Extracting contract elements	Chalkidis et al. (2017)	EN	MLL
19	Improving Sentence Retrieval from Case Law for Statutory Interpretation	Savelka et al. (2019)	EN	LTA
20	JEC-QA: a legal-domain question answering dataset	Zhong et al. (2020)	CN	PLI
21	LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts	Tuggener et al. (2020)	EN	CIT
22	MAPS: Scaling Privacy Compliance Analysis to a Million Apps	Zimmeck et al. (2019)	EN	CIT
23	Mining Legal Arguments in Court Decisions - Data and software (European Court of Human Rights (ECHR))	Habernal et al. (2022)	EN	ALD

Table 2 (continued)

ID	Title	Author(s)	Lang	Source
24	NLP for Consumer Protection: Battling Illegal Clauses in German Terms and Con-ditions in Online Shopping	Braun and Matthes (2021)	DE	CIT
25	Plain English Summarization of Contracts	Manor and Li (2019)	EN	PLI, MLL
26	Segmenting US Court Decisions into Functional and Issue Specific Parts	Šavelka and Ashley (2018)	EN LTA	MLL
27	The Creation and Analysis of a Website Privacy Policy Corpus	Wilson et al. (2016)	EN	CIT
28	The HOLJ Corpus. Supporting Summarisation of Legal Texts	Grover et al. (2004)	EN	PLI
29	Toward Domain-Guided Controllable Summarization of Privacy Policies	Keymanesh et al. (2020)	EN	MLL

(2021) consists of parallel documents in four languages and was only annotated in English and the annotations were then automatically transferred to other language versions of the documents.

5 Annotation process analysis

For each of the 29 data sets shown in Table 2, we analysed the annotation processes that lead to the labelled data set. We specifically focused on aspects related to the handling of disagreement within the process. In order to provide a meaningful comparison, we identified seven particularly relevant aspects that can be distinguished:

1. **Amount and type of documents** Legal data sets can contain a wide variety of document types, like court decisions, contracts or pairs of questions and answers. The type and amount of documents that are to be annotated have an impact on the annotation process and especially the type influences the amount of disagreement that is to be expected.
2. **Type of annotations** The type of annotations that can be made are even more diverse than the type of documents. Annotations can be made on different levels (from the document level to individual words) and with regard to different aspects, like the semantic role of a sentence or a legal assessment. Some types of annotations are more prone to disagreement than others.
3. **Pool of annotators, reviewers, and arbiters (amount and expertise)** The pool of annotators describes the people involved in the annotation process. We specifically look at the amount as well as their expertise. In the following, we will differentiate between three roles with relevance with regard to the handling of disagreement: We will speak of *annotators*, only if a person independently labels data, i.e. without having access to previous annotations by somebody else. We will speak of *reviewers* if a person makes annotations based on existing labels by one or multiple previous annotators. Finally, we will speak of *arbiters* for people that will be involved in the annotations process to resolve a disagreement between annotators. Arbiters therefore only annotate records that have disagreeing previous annotations. With regard to expertise and for the benefit of brevity, we will summarise legal scholars, lawyers, and other groups with special legal expertise under the category “experts”.
4. **Annotators per record** A large pool of annotators does not automatically imply that each record is also annotated by multiple persons. Therefore, we separately describe for each data set how many people independently annotated each record. For our analysis, we are particularly interested in data sets with more than one annotator per record.
5. **Reviewers per record** Similarly, we also provide the number of reviewers per record. Even if a record is only annotated by one person, disagreement can still arise if it is subsequently reviewed and the reviewer disagrees with the assessment of the annotator.

6. **Agreement metric** Inter-annotator agreement is widely seen as an indication of the difficulty of an annotation task, as well as the quality of annotations (Artstein 2017). Different metrics exist to compute the agreement between annotators. Such metrics can only be calculated if more than one annotator per record exists. Otherwise, this property is marked as not applicable (n.a.).
7. **Strategy for disagreement** The main focus of the analysis was put on how disagreement between annotators was handled during the annotation process. For corpora in which each record was only annotated by one person and no reviewers were involved, this property is not applicable (n.a.). In the analysis, we identified four repeating strategies that are applied to handle disagreement. These strategies are described in Sect. 5.3.

In the assessment, only the final round of annotations, which produces the labels for the corpus, was considered. Some of the data sets, e.g. by Roegiest et al. (2018), Hendrycks et al. (2021), and Chalkidis et al. (2017), were built using an iterative process, in which a first set of test documents was annotated jointly or discussed between annotators to refine the annotation guidelines. These pre-annotations are not reflected in our analysis. Some corpora also consist of an automatically labelled part and a manually labelled part, e.g. from Tuggener et al. (2020) and Zhong et al. (2020). The analysis only considers the manually labelled part. Finally, some corpora, e.g. from Šavelka and Ashley (2018) and Grover et al. (2004), had a small fraction of the overall corpus (usually around 10%) annotated by two annotators, in order to calculate inter-annotator agreement, while the rest of the corpus was only annotated by one person. In the analysis, we consider these cases as annotated by one annotator, since the majority of the corpus was created in this way. However, we will still provide the agreement metric that was used in the overview.

The results of the analysis are shown in Table 3. Missing information is indicated with a “–”, categories that are not applicable for the specific annotation process, e.g. the strategy to resolve a disagreement in cases where each item is annotated by just one annotator, are indicated with “n.a.” (not applicable). The overview shows the diversity of the analysed data sets: from data sets with just ten documents to data sets with tens of thousands of documents, a wide range is covered. At the same time, there are a lot of similarities between the different corpora. Out of the 29 data sets, ten consist of judgements, five of Terms and Conditions or Terms of Services, three of privacy policies, three of pairs of question and answers, two of contracts, and just one of the legislative texts. In the following sections, we describe patterns we identified that apply to a large number of the analysed data sets.

5.1 Lack of information

To our surprise, 13 out of the 29 publications did not provide all the information we analysed. Even arguably one of the most basic information, if a record was annotated by a single annotator or multiple annotators, was not provided by five out of 29 publications ($\sim$ 17%). Although, a number of different schemata have emerged for the description of data sets, e.g. from Gebru et al. (2021) and Holland et al. (2020), only

Table 3 Results of the annotation process analysis (IDs matching Table 2; "–" indicates missing information; "n.a." indicates categories that are not applicable to the specific data set)

ID	Docs	Annotations	Pool	Anno./item	Rev./item	Agree. metric	Disagree. strategy
1	10 judgments	Law school annotations	3 students	3	0	Pair-wise P, R, and F1	Majroity vote and expert arbiter
2	100 ToS	Unfair clauses	2–3 people	-	0	Cohen's kappa	–
3	4412 contracts	Relevance experts	Students	1	1	-	Expert reviewer
4	32 laws	Relevance	6 experts	1	0	n.a	n.a
5	2572 judgments	Relevance	3 experts	-	0	n.a	–
6	75 judgments	Polarity of sentences	2 annotators with advanced degrees in english and 1 legal experts	1	1	–	Expert reviewer
7	277 judgments	Rhetoerical roles	Law students	–	0	–	–
8	8964 judgments	Similarity	Legal experts	3	0	–	–
9	10,000 judgments	Q& A pairs	Legal experts	–	–	–	–
10	Laws	Semantic types	3 experts	2	0	–	Arbiter
11	100 ToS	Unfair clauses	2–3 people	-	0	Cohen's kappa	–
12	172 T &C	Clause topics	1 student authors 1, expert	2	0	% of agreement	Expert arbiter
13	Contract and reports	Clause types	Experts	2	1	Soft F1	Reviewer
14	354 judgments	Rhetorical roles	35 students	3	0	Fleiss' kappa	Majority vote and expert arbiter
15	500 contracts	Clause types experts	Students	4	0	–	Expert arbiter
16	200 judgments	Argument structure	Expert	1	0	n.a	n.a
17	42 judgments	Argument structure	2 experts	2	1	Cohen's kappa	Arbiter
18	993 contracts	Clause types	10 students	1	0	n.a	n.a
19	4635 sentences from judgments	Meaning of statuatory phrases	3 people	2	1	Krippen–Dorff's alpha	Reviewer

Table 3 (continued)

ID	Docs	Annotations	Pool	Anno./item	Rev./item	Agree. metric	Disagree. strategy
20	377 questions	Q & A pairs	–	–	–	–	–
21	NDAs	Clause types	3 experts	–	–	–	–
22	350 privacy policies	Data practices	1 expert	1	0	Krippen–Dorff's alpha	n.a
23	373 judgments	Argument structure	6 students, 2 experts	>1	>0	Krippen–Dorff's alpha	Expert arbiter
24	1186 T&C clauses	Void clauses	5 experts	2	0	% of agreement	Forced agreement
25	421 clauses ToS	Simplified summaries	Crowd, 1 expert	–	1	% of agreement	Expert reviewer
26	316 judgments	Functional and issue specific parts	2 authors	1	0	Adapted accuracy	n.a
27	115 privacy policies	Data practices	10 students	3		Fleiss' kappa	Vote with threshold
28	40 judgments	Rhetorical roles	2 people	1		Cohen's kappa	n.a
29	151 privacy policies	Privacy risks	Crowd, 1 expert	–	1	–	Expert reviewer

one of the analysed data sets followed such a schema to provide a structured representation of the data set introduced. For the 17 data sets that explicitly disclosed that each record was annotated by more than one person, only 12 provide a measurement for the agreement between annotators, although inter-annotator agreement is an important metric for the validity of the annotation process (Artstein and Poesio 2008). For the five data sets that did not specify how many annotators labelled each item, none provides an agreement measurement, possibly hinting at not having used multiple annotators per record.

Without some of the basic information, it is very difficult to judge the quality of a given data set by anyone who might want to use it in their research. We, therefore, believe that such information should be included in each publication introducing a new data set.

5.2 Pool of annotators, reviewers, and arbiters

Except for two data sets that used existing annotations from a crowd-sourcing website (Manor and Li 2019; Keymanesh et al. 2020), all data sets were solely annotated by domain experts, either students of law or people with a law degree. Based on the scope of our analysis, which specifically only includes data sets with annotations that require legal knowledge, this was to be expected. Nine out of 29 data sets use students to some extent in their annotation process. However, only four of them rely solely on student annotators, without higher-qualified expert reviewers or arbiters. Given the high expertise of the people involved in the annotation, it is not surprising that most data sets were labelled by a small pool of annotators. 17 out of 29 data sets have a pool of three people or less. Given the small number of available people overall, it is also not surprising that in most data sets, items are not annotated by more than 3 people. In at least five data sets, each item was only annotated by one person.

5.3 Strategies for disagreement

The focus of the analysis is on the disagreement between annotators and how it is handled in the annotation process. That includes

- how disagreement between annotators is reported in the publications (i.e. which metrics are used to calculate inter-annotator agreement),
- the way a gold standard is derived from multiple, possibly conflicting, annotations, and
- the way possible disagreement is represented in the final data sets.

For the representation in the final data set, it has to be concluded that none of the data sets investigated represents disagreement during the annotation process in any form in their final data sets. All corpora only contain the final “gold standard” annotations.

With regard to the publications and the report of inter-annotator agreement, we found seven different metrics used in the 29 data sets we investigated, including

Table 4 Avg. agreement score within the analysed data sets for the three most frequently used metrics

Agreement metric	Avg. score
Cohen's kappa	0.76
Fleiss' kappa	0.675
Krippendorff's alpha	0.677

standardised metrics like Cohen's kappa (Cohen 1968), Fleiss' kappa (Fleiss 1971), and Krippendorff's alpha (Krippendorff 2018), but also some tailor-made or modified metrics. This diversity of metrics limits the comparability and, especially in the case of the tailored metrics, the interpretability of the reported inter-annotator agreement.

For the three standard metrics, Table 4 shows the average scores reported in the analysed data sets. For Cohen's and Fleiss' kappa, the values of 0.76 and 0.68 respectively can be interpreted, according to Landis and Koch (1977), as a "substantial" agreement, although on the lower end of the range between 0.61 and 0.80. For Krippendorf's alpha, (Krippendorff 2018, p. 241) himself says that "it is customary to require $\alpha \geq .800$. Where tentative conclusions are still acceptable, $\alpha \geq .667$ is the lowest conceivable limit". Of the three analysed data sets that report Krippendorff's alpha, the highest reported value is 0.78. These low inter-annotator agreement values for tasks that are, by legal standards, not among the most controversial, like argument structure annotation or annotation of data practices emphasise the need to reflect on the practices around the handling of disagreement during the annotation of legal data sets.

With regard to the practices that are applied to derive a gold standard from disagreeing annotations, the strategies found can be classified into four categories, which are explained in detail in the following sections. Such strategies are only applicable in cases where more than one person is involved in the annotation of each item, otherwise, the annotation made by the (sole) annotator is automatically used as gold standard.

5.3.1 (Majority) vote

One of the simplest approaches to dissolve disagreement between annotators to get to a gold standard is a majority vote: Each independently made annotation represents one vote and the annotation which gets the largest number of votes is used as gold standard. This strategy is often applied for crowd-sourced annotation, i.e. if there is a large number of independent annotators. For small numbers of annotators, it is hard to see how 2 vs 1, or 3 vs 2 annotations could be a decisive difference that can be trusted. Given that all the analysed data sets use a small number of annotators, it is not surprising, that none of the data sets uses a pure majority vote strategy. However, the data set from Wyner et al. (2013) did use a majority vote strategy for a subset of their label categories, which they deemed to be "simple annotations" based on domain knowledge. A (pure) majority vote can only be used in cases where each item is annotated by an uneven number of annotators, because, otherwise, we could end up with a tie.

Wilson et al. (2016) use a voting approach which uses a threshold (below the majority) and therefore also allows labels in the gold standard that were assigned by only a minority of the annotators. They justify this practice with the assumption that “data labeled by only one or two skilled annotators also have substantial value and merit retention” (Wilson et al. 2016). However, they specifically exclude disagreement from this practice and only add minority labels if they are supplementary to the majority decision, e.g. by adding a label that further specifies an already existing majority label or is of the same category.

5.3.2 Forced agreement

Another strategy we identified to solve disagreement is what we call “forced agreement”: If annotators disagree, they review the conflicting annotations together and have to deliver a shared final annotation. Such an approach is not feasible for a larger number of annotators and in general, seems only applicable to areas where a strong notion of one (exclusively) correct answer exists. In cases with more room for interpretation, it could be that the annotators cannot agree, leaving us without a gold standard.

5.3.3 (Expert) reviewer

Six out of the 29 data sets used a strategy in which a reviewer assigns the final gold standard label based on annotations made by one or multiple annotators. In four out of six cases, the reviewer has higher formal expertise than the original annotators, making them an expert reviewer. In cases where items are annotated by just one annotator and reviewed by one reviewer, arguably the value compared to just having the reviewer making the annotations in the first place is relatively low because the gold standard is still completely dependent on the assessment of just one person, the reviewer. In cases where the annotation process is laborious, e.g. if subsections of a text have to be found and marked, this method can bring cost benefits when using a higher-qualified reviewer.

If the reviewer receives annotations from more than one annotator per item, there are different possible strategies for how the final gold standard is decided. In some cases, the reviewer has full autonomy and can (in theory) overrule any annotations, even if all annotators agreed on a label. In other cases, the reviewer can only choose between annotations suggested by the original annotators. Often, however, no detailed information is provided as to how the reviewer comes to the final gold standard and whether or not they are able to overrule the work of the annotators.

5.3.4 (Expert) arbiter

The arbiter strategy is very similar to the reviewer strategy, however, arbiters are only consulted in case of disagreement. Therefore, this strategy can only be applied if each item is annotated by more than one person and disagreement can arise. If there is consensus about a label between all annotators, it is automatically added to the gold standard. If there is disagreement, the arbiter gets to decide on the final label. As for the reviewer strategy, it can again be differentiated between strategies where the reviewer has to choose between the labels suggested by annotators and strategies where they can choose the final label freely. Seven out of 29 data sets used this strategy. In five of those cases, the arbiter had higher formal expertise, making them an expert arbiter.

6 Suggestions

Based on the analysis of existing legal data sets and their annotation processes described in the previous section and the assumption that the representation of disagreement between annotators in legal data sets can be beneficial, as we have argued in the introduction, we derived suggestions on how to handle disagreement in the annotation of legal data sets. The suggestions concern the annotation process itself, the description of the process, and the reporting of the outcome of this process in the data itself and the accompanying publication.

All suggestions can be implemented independently of each other. Even if a data set is annotated through a simple majority vote process, e.g. for cost and efficiency reasons, reporting in more detail on the annotation process, as suggested in Sect. 6.2, can increase the possibilities for re-use of the data set and creates more transparency. Additionally, if items in a data set are annotated by more than one person, the information about disagreement is already available and just has to be published alongside the derived gold standard, for which no significant additional effort is necessary. The main goal of the suggested annotation process described in Sect. 6.1 is to ensure that all disagreement that is present in the data set is based on genuine disagreement, rather than a lack of attention.

6.1 Annotation process

Our suggestions for the annotation process are based on the fact that expert annotators in the legal domain are expensive and it is therefore unrealistic to propose processes that involve a large number of annotators. At the same time, the suggestions are based on the assumption that, for a reliable annotation, no label should be based on the decision of just one annotator. Therefore, we suggest having every item annotated by at least two independent annotators with a similar level of expertise. Because of the small number of annotators, we suggest using an arbiter to resolve disagreements between annotators, rather than a voting approach. Using two annotators and one arbiter, rather than a majority vote of three annotators, guarantees that a final decision is reached: For non-binary classifications, three annotators could suggest three different labels. The arbiter, on the other hand, will have to choose between, at most, two options provided by the annotators, ensuring that a final label for the gold standard is picked.

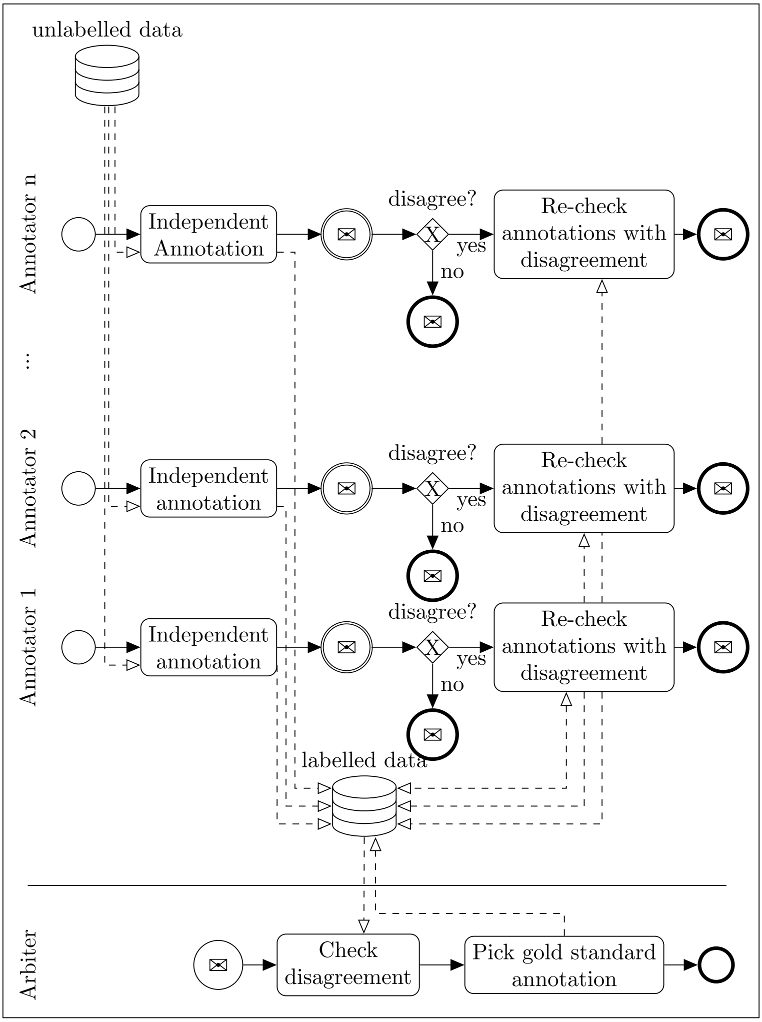

Fig. 1 Suggested annotation process

Because the desired outcome of the suggested annotation process is a corpus that does not just contain a gold standard but also possible disagreements between annotators, we suggest an additional feedback loop to make sure that all disagreeing are based on actual disagreement, rather than lack of attention from one of the annotators. The process we suggest is sketched in Fig. 1 using the Business Process Model and Notation (BPMN) (Chinosi and Trombetta 2012). From BPMN, we use three core elements for the process model: tasks (represented by squares with rounded corners), events (represented by circles), and gateways (represented by diamonds). Circles with thin outlines represent start events, circles with thick outlines represent end events, and circles with double outlines represent intermediate events where the process is temporarily suspended. Envelopes in circles indicate messaging between different parallel activities.

In the process presented in Fig. 1, each annotator independently starts to annotate their assigned items. After finishing their individual annotations, annotators wait for all annotators to finish this process. Items for which disagreeing annotations exist are then presented again to the original annotators. Annotators can only see their own annotations and are asked to double-check whether the annotation is correct. The goal of this process is *not* to achieve consensus, but to make sure that the disagreement originates from a disagreement in the subject matter and not a mistake. The result is a labelled corpus that contains all annotations from all annotators. Once the annotators are finished, a message is sent to the arbiter.

In order to derive a gold standard from these annotations, an arbiter revisits all items with disagreeing annotations and chooses one of the given annotations the be the gold standard. This happens after the arbiter received a message from all annotators that they finished the annotation process. In this way, it is ensured that each gold standard label is at least supported by two people in the annotation process. The final corpus then contains the labels from the independent annotation process as well as the gold standard (see also Sect. 6.2).

The focus of this article is on making disagreement explicit, rather than ways of resolving disagreement in order to derive a gold standard. Nevertheless, it should be mentioned that, especially in cases with a larger number of annotators, it might be useful to introduce additional constraints on the arbiter, especially if the arbiter has the same level of expertise as the annotators. If, for example, nine annotators choose label A and one annotator chooses label B, we might want to restrict the arbiter by defining certain thresholds a label has to achieve in order to be selectable by the arbiter.

6.2 Reporting

Publications that describe newly introduced data sets should provide sufficient information for others to assess whether the data set is appropriate for their own use. Following documentation standards like “Datasheets for Datasets” from Gebru et al. (2021) can help to make sure all relevant information is provided. However, the standard is focused on describing large data sets and the acquisition of “raw” (i.e. unlabelled) data. The annotation process is only a side note. To provide transparency about the annotation process, especially with regard to disagreement, we suggest including at least the following information:

- pool of people involved in the annotation (number and expertise),
- number of independent annotators per item,
- number and expertise of reviewers/arbiters,
- inter-annotator agreement using a standardised metric (where applicable), and
- detailed description of the annotation process including handling of disagreement.

In the data set itself, we suggest including all original annotations, where possible, after removing non-intentional disagreement, alongside the gold standard. In cases where there are no concerns with regard to the privacy of the annotators or other issues, we would also suggest providing the individual annotations in a way that makes it transparent which annotations have been made by the same annotator.

7 Conclusion

In this article, we analyse how disagreement is handled in the annotation of legal data sets. We identified 29 legal data sets out of a list of more than 120 data sets that have undergone an annotation process that is relevant to the question. The analysis shows that in all 29 cases, the final artefact (i.e. the annotated data set) does not contain any hint whatsoever of disagreement that might have occurred during the annotation process. The analysis also shows that many of the publications introducing new data sets are lacking relevant information making it effectively impossible for other scientists to judge the reliability of the labels within the data.

We hope that this article can spark a discussion within the community about why we see it as normal that data sets present a single “truth” where experts disagree. The suggestions presented in this article can provide first guidance on how disagreement between annotators can be made more transparent in the future. Even if this information cannot yet be productively used with most predictive models, conserving and providing this information cannot only help us to understand the subject of the annotation better but in the future also might be of value for models that are able to use this information to provide more nuanced and balanced predictions than the current state-of-the-art technologies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10506-023-09369-4.

References

Akhtar S, Basile V, Patti V (2020) Modeling annotator perspective and polarized opinions to improve hate speech detection. In: Proceedings of the AAAI conference on human computation and crowdsourcing, vol 8, no 1, pp 151–154. https://doi.org/10.1609/hcomp.v8i1.7473

Artstein R (2017) Inter-annotator agreement. Springer, Dordrecht, pp 297–313. https://doi.org/10.1007/978-94-024-0881-2_11

Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. Comput Linguist 34(4):555–596. https://doi.org/10.1162/coli.07-034-R2

Basile V, Cabitza F, Campagner A et al. (2021) Toward a perspectivist turn in ground truthing for predictive computing. CoRR arxiv:2109.04270

Beigman Klebanov B, Beigman E, Diermeier D (2008) Analyzing disagreements. In: Coling 2008: proceedings of the workshop on human judgements in computational linguistics. Coling 2008 Organizing Committee, Manchester, UK, pp 2–7. https://aclanthology.org/W08-1202

Borchmann Ł, Wisniewski D, Gretkowski A et al. (2020) Contract discovery: Dataset and a few-shot semantic retrieval challenge with competitive baselines. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 4254–4268. https://doi.org/10.18653/v1/2020.findings-emnlp.380

Braun D, Matthes F (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. Association for Computational Linguistics, Online, pp 93–99. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.10

Braun D, Matthes F (2022) Clause topic classification in German and English standard form contracts. In: Proceedings of the fifth workshop on e-commerce and NLP (ECNLP 5). Association for Computational Linguistics, Dublin, Ireland, pp 199–209. https://doi.org/10.18653/v1/2022.ecnlp-1.23

Campagner A, Ciucci D, Svensson CM et al. (2021) Ground truthing from multi-rater labeling with three-way decision and possibility theory. Inf Sci 545:771–790. https://doi.org/10.1016/j.ins.2020.09.049

Chalkidis I, Androutsopoulos I, Michos A (2017) Extracting contract elements. In: Proceedings of the 16th edition of the international conference on artificial intelligence and law. Association for Computing Machinery, New York, NY, USA, ICAIL '17, pp 19–28. https://doi.org/10.1145/3086512.3086515

Chalkidis I, Jana A, Hartung D et al. (2022) LexGLUE: a benchmark dataset for legal language understanding in English. In: Proceedings of the 60th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Dublin, Ireland, pp 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

Chan B, Schweter S, Möller T (2020) German's next language model. In: Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598

Chinosi M, Trombetta A (2012) BPMN: an introduction to the standard. Comput Stand Interfaces 34(1):124–134. https://doi.org/10.1016/j.csi.2011.06.002

Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213

Davani AM, Díaz M, Prabhakaran V (2022) Dealing with disagreements: looking beyond the majority vote in subjective annotations. Trans Assoc Comput Linguist 10:92–110. https://doi.org/10.1162/tacl_a_00449

Drawzeski K, Galassi A, Jablonowska A et al. (2021) A corpus for multilingual analysis of online terms of service. In: Proceedings of the natural legal language processing workshop 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 1–8. https://doi.org/10.18653/v1/2021.nllp-1.1

Duan X, Wang B, Wang Z et al. (2019) CJRC: a reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. In: Sun M, Huang X, Ji H et al. (eds) Chinese computational linguistics. Springer, Cham, pp 439–451

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

Gebru T, Morgenstern J, Vecchione B et al. (2021) Datasheets for datasets. Commun ACM 64(12):86–92. https://doi.org/10.1145/3458723

Glaser I, Scepankova E, Matthes F (2018) Classifying semantic types of legal sentences: portability of machine learning models. In: Legal knowledge and information systems. IOS Press, pp 61–70

Gonzalez D, Zimmermann T, Nagappan N (2020) The state of the ML-universe: 10 years of artificial intelligence & machine learning software development on GitHub. In: Proceedings of the 17th international conference on mining software repositories. Association for Computing Machinery, New York, NY, USA, MSR '20, pp 431–442. https://doi.org/10.1145/3379597.3387473

Grover C, Hachey B, Hughson I (2004) The HOLJ corpus. Supporting summarisation of legal texts. In: Proceedings of the 5th international workshop on linguistically interpreted Corpora. COLING, Geneva, Switzerland, pp 47–54. https://aclanthology.org/W04-1907

Guha N (2021) Datasets for machine learning in law. Tech. rep., Stanford University, https://github.com/neelguha/legal-ml-datasets

Habernal I, Faber D, Recchia N et al. (2022) Mining legal arguments in court decisions. arXiv preprint https://doi.org/10.48550/arXiv.2208.06178

Hendrycks D, Burns C, Chen A et al. (2021) CUAD: an expert-annotated NLP dataset for legal contract review. CoRR arxiv:2103.06268

Holland S, Hosny A, Newman S et al. (2020) The dataset nutrition label. Data protection and privacy, volume 12: data protection and democracy 12:1

Jamison E, Gurevych I (2015) Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, Portugal, pp 291–297. https://doi.org/10.18653/v1/D15-1035

Kalamkar P, Tiwari A, Agarwal A et al. (2022) Corpus for automatic structuring of legal documents. CoRR arxiv:2201.13125

Keymanesh M, Elsner M, Sarthasarathy S (2020) Toward domain-guided controllable summarization of privacy policies. In: NLLP@ KDD, pp 18–24

Klemen M, Robnik-Šikonja M (2022) ULFRI at SemEval-2022 task 4: leveraging uncertainty and additional knowledge for patronizing and condescending language detection. In: Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022). Association for Computational Linguistics, Seattle, United States, pp 525–532. https://doi.org/10.18653/v1/2022.semeval-1.73

Kralj Novak P, Scantamburlo T, Pelicon A et al. (2022) Handling disagreement in hate speech modelling. In: Ciucci D, Couso I, Medina J et al. (eds) Information processing and management of uncertainty in knowledge-based systems. Springer, Cham, pp 681–695

Krippendorff K (2018) Content analysis: an introduction to its methodology, 4th edn. Sage Publications, Thousand Oaks

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Li S (2017) A corpus-based study of vague language in legislative texts: strategic use of vague terms. Engl Specif Purp 45:98–109. https://doi.org/10.1016/j.esp.2016.10.001

Lippi M, Pałka P, Contissa G et al. (2019) Claudette: an automated detector of potentially unfair clauses in online terms of service. Artif Intell Law 27(2):117–139

Locke D, Zuccon G (2018) A test collection for evaluating legal case law search. In: The 41st international ACM SIGIR conference on research & development in information retrieval. Association for Computing Machinery, New York, NY, USA, SIGIR '18, pp 1261–1264. https://doi.org/10.1145/3209978.3210161

Louis A, Spanakis G (2022) A statutory article retrieval dataset in French. In: Proceedings of the 60th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Dublin, Ireland, pp 6789–6803. https://doi.org/10.18653/v1/2022.acl-long.468

Lübbe-Wolff G (2022) Beratungskulturen: Wie verfassungsgerichte arbeiten, und wovon es abhängt, ob sie integrieren oder polarisieren. Tech. rep, Konrad-Adenauer-Stiftung

Manor L, Li JJ (2019) Plain English summarization of contracts. In: Proceedings of the natural legal language processing workshop 2019. Association for Computational Linguistics, Minneapolis, Minnesota, pp 1–11. https://doi.org/10.18653/v1/W19-2201, https://aclanthology.org/W19-2201

Ostendorff M, Blume T, Ostendorff S (2020) Towards an open platform for legal information. In: Proceedings of the ACM/IEEE joint conference on digital libraries in 2020. Association for Computing Machinery, New York, NY, USA, JCDL '20, pp 385–388. https://doi.org/10.1145/3383583.3398616

Ovesdotter Alm C (2011) Subjective natural language problems: motivations, applications, characterizations, and implications. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, pp 107–112. https://aclanthology.org/P11-2019

Poudyal P, Savelka J, Ieven A et al. (2020) ECHR: legal corpus for argument mining. In: Proceedings of the 7th workshop on argument mining. Association for Computational Linguistics, Online, pp 67–75. https://aclanthology.org/2020.argmining-1.8

Prabhakaran V, Mostafazadeh Davani A, Diaz M (2021) On releasing annotator-level labels and information in datasets. In: Proceedings of the Joint 15th linguistic annotation workshop (LAW) and 3rd designing meaning representations (DMR) workshop. Association for Computational

Linguistics, Punta Cana, Dominican Republic, pp 133–138. https://doi.org/10.18653/v1/2021.law-1.14

Ramponi A, Leonardelli E (2022) DH-FBK at SemEval-2022 task 4: Leveraging annotators' disagreement and multiple data views for patronizing language detection. In: Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022). Association for Computational Linguistics, Seattle, United States, pp 324–334. https://doi.org/10.18653/v1/2022.semeval-1.42

Roegiest A, Hudek AK, McNulty A (2018) A dataset and an examination of identifying passages for due diligence. In: The 41st international ACM SIGIR conference on research & development in information retrieval. Association for Computing Machinery, New York, NY, USA, SIGIR '18, pp 465–474. https://doi.org/10.1145/3209978.3210015

Rottger P, Vidgen B, Hovy D et al. (2022) Two contrasting data annotation paradigms for subjective NLP tasks. In: Proceedings of the 2022 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 175–190. https://doi.org/10.18653/v1/2022.naacl-main.13

Sachdeva P, Barreto R, Bacon G et al. (2022) The measuring hate speech corpus: leveraging Rasch measurement theory for data perspectivism. In: Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022. European Language Resources Association, Marseille, France, pp 83–94. https://aclanthology.org/2022.nlperspectives-1.11

Sas C, Capiluppi A (2022) Antipatterns in software classification taxonomies. J Syst Softw 190(111):343. https://doi.org/10.1016/j.jss.2022.111343

Šavelka J, Ashley KD (2018) Segmenting us court decisions into functional and issue specific parts. In: Legal knowledge and information systems. IOS Press, pp 111–120

Savelka J, Xu H, Ashley KD (2019) Improving sentence retrieval from case law for statutory interpretation. In: Proceedings of the seventeenth international conference on artificial intelligence and law. Association for Computing Machinery, New York, NY, USA, ICAIL '19, pp 113–122. https://doi.org/10.1145/3322640.3326736

Schwarzer M (2022) awesome-legal-data. Tech. rep., Open Justive e.V., https://github.com/openlegaldata/awesome-legal-data

Steinberger R, Pouliquen B, Widiger A et al. (2006) The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the fifth international conference on language resources and evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf

Sudre CH, Anson BG, Ingala S et al. (2019) Let's agree to disagree: learning highly debatable multirater labelling. In: Shen D, Liu T, Peters TM et al. (eds) Medical image computing and computer assisted intervention—MICCAI 2019. Springer, Cham, pp 665–673

Tiwari A, Kalamkar P, Agarwal A et al. (2022) Must-read papers on legal intelligence. Tech. rep., OpenNyAI. https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline

Tuggener D, von Däniken P, Peetz T et al. (2020) LEDGAR: a large-scale multi-label corpus for text classification of legal provisions in contracts. In: Proceedings of the twelfth language resources and evaluation conference. European Language Resources Association, Marseille, France, pp 1235–1241. https://aclanthology.org/2020.lrec-1.155

Urchs S, Mitrović J, Granitzer M (2021) Design and implementation of German legal decision corpora. In: Proceedings of the 13th international conference on agents and artificial intelligence—volume 2: ICAART, INSTICC. SciTePress, pp 515–521. https://doi.org/10.5220/0010187305150521

Walker VR, Strong SR, Walker VE (2020) Automating the classification of finding sentences for linguistic polarity. In: Proceedings of the fourth workshop on automated semantic analysis of information in legal text

Waltl B (2022) Legal text analytics. Tech. rep., Liquid Legal Institute e.V. https://github.com/Liquid-Legal-Institute/Legal-Text-Analytics

Wilson S, Schaub F, Dara AA et al. (2016) The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, Germany, pp 1330–1340. https://doi.org/10.18653/v1/P16-1126, https://aclanthology.org/P16-1126

Wu Y, Wang N, Kropczynski J et al. (2017) The appropriation of GitHub for curation. PeerJ Comput Sci 3:e134

Wyner A, Peters W, Katz D (2013) A case study on legal case annotation. In: Legal knowledge and information systems. IOS Press, pp165–174

Xiao C, Zhong H, Guo Z et al. (2019) CAIL2019-SCM: a dataset of similar case matching in legal domain. CoRR arxiv:1911.08962

Xiao C, Zhong H, Sun Y (2021) Must-read papers on legal intelligence. Tech. rep., Tsinghua University. https://github.com/thunlp/LegalPapers

Zahidi Y, El Younoussi Y, Azroumahli C (2019) Comparative study of the most useful Arabic-supporting natural language processing and deep learning libraries. In: 2019 5th international conference on optimization and applications (ICOA), pp 1–10. https://doi.org/10.1109/ICOA.2019.8727617

Zhong H, Xiao C, Tu C et al. (2020) JEC-QA: a legal-domain question answering dataset. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no. 05, pp 9701–9708. https://doi.org/10.1609/aaai.v34i05.6519

Zimmeck S, Story P, Smullen D et al. (2019) Maps: scaling privacy compliance analysis to a million apps. Proc Priv Enhanc Technol 2019:66

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.