A Multilingual Approach to Identify and Classify Exceptional Measures Against COVID-19

Georgios Tziafas[1], **Eugenie de Saint-Phalle**[2], **Wietse de Vries**[3], **Clara Egger**[2] and **Tommaso Caselli**[3]
1. Artificial Intelligence 2. IRIO 3. CLGC
University of Groningen
g.tziafas@student.rug.nl,
{e.de.saint-phalle|wietse.de.vries}@rug.nl
{c.m.egger|t.caselli}@rug.nl

Abstract

The COVID-19 pandemic has witnessed the implementations of exceptional measures by governments across the world to counteract its impact. This work presents the initial results of an on-going project, EXCEPTIUS, aiming to automatically identify, classify and compare exceptional measures against COVID-19 across 32 countries in Europe. To this goal, we created a corpus of legal documents with sentence-level annotations of eight different classes of exceptional measures that are implemented across these countries. We evaluated multiple multi-label classifiers on a manually annotated corpus at sentence level. The XLM-RoBERTa model achieves highest performance on this multilingual multi-label classification task, with a macro-average F1 score of 59.8%.

1 Introduction

The increasing availability of digitized and publicly available legal documents has boosted political scientists, legal scholars, lawyers and policy-makers to apply Human Language Technologies (HLT) to discover, analyze, digest, and use automatically extracted information. All of these operations fall under the larger area of study that can be labeled as Legal Artificial Intelligence, or LegalAI (Zhong et al., 2020). Similarly to any domain that wants to apply HLT to process written documents, LegalAI requires the development of both domain-specific languages resources (e.g., annotated corpora or embedding representations (Chalkidis et al., 2019c; Kornilova and Eidelman, 2019; Holzenberger et al., 2020; Luz de Araujo et al., 2020; Samy et al., 2020)) and tools (e.g., language specific natural language processing pipelines (Koeva et al., 2020; Moreno-Schneider et al., 2020) or pretrained language models (Chalkidis et al., 2020)). At the same time, LegalAI presents an additional challenge when it comes to the development of multilingual systems. With very few exceptions (e.g., EU-level legislation), in the domain of LegalAI, natural languages are strictly connected to countries, meaning that different legislative systems and practices may be in place. For instance, although French-speaking countries share a common language, their legal traditions widely differ making comparisons across legal systems uneasy. Nevertheless, the adoption of a multilingual approach may prove valuable especially to legal practitioners and scholars as well as political scientists and policy-makers who are increasingly interested in comparing legal systems and examining how legal concepts "travel" across time and spaces. Commonly used data collection and analysis techniques - largely relying on manual or rule-based coding of legal documents - have so far prevented the development of meaningful and systematic analyses allowing the comparison of fine-grained classes.

In this paper we investigate the potential of multilingual pretrained language models in order to facilitate the analysis, exploration, and comparison of legal texts on COVID-19 exceptional measures. Our major contributions can be summarised as follows:

- the creation of a **new corpus** of legislative documents from 21 European countries manually annotated for exceptional measures against COVID-19 (Sections 2 and 3);
- the development of a **rich taxonomy** (eight classes and 83 subclasses) to identify and compare exceptional measures in a consistent way (Sections 4);
- the development of a **multi-label classifier** based on XLM-RoBERTa (Conneau et al., 2019) to identify exceptional measures at sentence level (Section 5).

2 The Exceptional Measures Against COVID-19 in Europe

The COVID-19 pandemic has led governments around the world to take exceptional measures

Natural Legal Language Processing Workshop 2021, pages 46–62
November 10, 2021. ©2021 Association for Computational Linguistics

in order to contain the spread of the virus. Such exceptional decision-making has seen executives challenge the scope and legality of their powers, as well as impose restrictions on democratic processes, the rule of law, fundamental rights and civil liberties. These exceptional measures considerably vary from one country to another, even in cases where some forms of coordination are claimed to be in place, like in the European Union. Countries sharing close political culture and institutions reacted in contrasting ways, as attested by the sharp difference between the Belgian and Dutch responses to the crisis between March and June 2020. While the Dutch government implemented one of the softest approaches in Europe relying on people's compliance with governmental recommendations, Belgium introduced very early a strict lockdown. This diversity is not only practical but also semantic. While many countries relied on a "lockdown" to contain the spread of the virus, restrictions vary in scope while enforcement modalities are unequally coercive (Engler et al., 2021; Egger et al., 2021). This fragmented political response sparked interest from researchers in political science, economics, and law that started to trace exceptional decision-making in times of COVID-19 (Porcher, 2020; Hale et al., 2021). To the best of our knowledge, all current data collection efforts are based on manual or rule-based methods applied to press releases (Hopkins and King, 2007; Grimmer and Stewart, 2013; Wiedemann, 2013; Wettstein, 2014) or on experts survey. Yet, in the specific case of the COVID-19, such methods suffer from three core limitations:

Decisions were taken on different legal bases There is a variety of legislative tools that have been put in action across countries to counteract the spreading of the virus. Some governments activated crisis-management instruments and legal frameworks, including, but not limited to, the activation of state of emergency provisions (Bjørnskov and Voigt, 2021) that predate the crisis. Others took decisions in an *ad hoc* manner on the basis of executive, legal or administrative acts taken not only at the national but also at the subnational level.

Measures evolved quickly At least for the first wave of the COVID-19 (late January - June 2020), measures evolved on a weekly, and sometimes, on a daily basis, requiring researchers to handle a very large amount of constantly evolving textual data. Since the application of close reading methods (i.e., extensive manual annotation) to such a large amount of documents is a daunting task to perform over a reasonable period of time, most of the competing research teams opted for collecting data on broad classes of events (lockdown, border closure, state of emergencies) based on press releases, conferences or experts opinions. The lack of fine grained classes derived from legal texts provides a false impression of homogeneity between various governmental responses, especially in a context of semantic ambiguity about the measures used.

Different countries, multiple languages Being a pandemic, the COVID-19 emergency affected the entire world. This global condition is accompanied by a rich and diverse language composition of any corpus created to investigate and compare the legislative measures of different countries. The intrinsic multi-lingual nature of this corpus has raised additional challenges for coding methods traditionally used in social sciences. Only a few legal texts are translated in English and some national languages are spoken by a fairly limited number of people.

Against such a background, we have initiated a research project, EXCEPTIUS (Exceptional measures in times of COVID-19) [1] to collect and document metrics of exceptionalism in 30 countries of the European Economic Area (EEA), plus UK and Switzerland, starting from late January 2020. EXCEPTIUS intends to address the abovementioned challenges in the analysis of COVID-19 measures in three ways. First, measures are automatically captured from a homogeneous corpus of legal sources uniquely allowing researchers to analyse the diversity of the legal instruments used to contain the COVID-19 pandemic. Press releases or expert surveys commonly used in competing projects only capture such dimensions indirectly and imperfectly. Second, our project defines the most comprehensive taxonomy of exceptional measures in the field of democratic governance, the rule of law and fundamental rights and liberties. The automatic application of such taxonomy to a comprehensive legal corpus allows to conciliate the need to rely on fine-grained categories with the constraints deriving from the analysis of a large and constantly evolving corpus. Last, we adopt a

[1] For a description of the research project and initial results, see https://exceptius.com/

multi-lingual approach to automatically analyse the sources of COVID-19 legislation, limiting the bias associated with the translations of the original texts in English.

The reliance on multi-lingual methods adopts a philosophical perspective of Artificial Intelligence (AI) as a problem-solving tool rather than as an adaptive mechanism mimicking human abilities (Winograd, 1997; Auernhammer, 2020; Caselli et al., 2021). The "intelligent" systems developed in this project (see Section 5) do not aim at substituting humans but are designed to account for different development cycles *with* humans in the loop. The use of automatic methods based on HLT allows us to overcome in a smart and fast way the three challenges previously described.

3 Corpus Collection

The corpus collection process has been overseen by four political science experts working in partnership with national legal experts. All documents were retrieved from official governmental websites that publish legal acts. The identification of the relevant documents has been done by means of 4 keywords (i.e., "COVID", "COVID-19", "Coronavirus" and "Health emergency"). For each language, the corresponding language specific keywords were used. In this initial phase, we focus on a sample of 19 EEA countries on measures adopted at the national level. To do so, we identify publicly available links to relevant documents [2] plus UK and Switzerland. We could not find corresponding documents for two countries of the EEA (i.e., Bulgaria and Greece). All documents have been collected either by manually downloading them or by automatic scraping.[3] For countries with more than one official language (e.g., Switzerland), legal acts were collected in all available languages.[4]

A total of 6,449 documents has been collected and stored in text format so far. Documents form a homogeneous set of existing COVID-19 legislation. Such legislation however includes a variety of texts adopting by political authorities acting at different levels. Beside legal acts - adopted by national parliaments - the corpus also includes executive acts - adopted by governmental authorities which were granted exceptional powers during the COVID-19 crisis - and administrative acts which mainly specifies how the implementation modalities of measures adopted by parliaments and governments. The distribution of the documents per country varies greatly: from 9 on the federal level in Germany to 969 in Slovenia, with 212.5 being the median. These differences are mainly due to a variability across the countries concerning the institutional levels responsible for taking actions against the spread of COVID-19: for example, the low number of German documents reflects the fact that the *Länder* were responsible for enacting COVID-related measures. We further process all documents using the SpaCy UDPipe 2 NLP pipeline[5] (Straka and Straková, 2017) using models trained on the Universal Dependencies project (Nivre et al., 2016, 2020). Although the UDPipe models may be sub-optimal for processing legal documents, the availability of models and unified representations (i.e., same sets of labels for parts-of-speech tagging and dependency relations) for all the languages of the corpus is an advantage. The SpaCy UDPipe 2 pipeline successfully processed 6,049 documents, providing sentence splitting, tokenization, part-of-speech tagging, and dependency relations. Not processed documents are due to issues in the text format due to the conversion process from pdf or other formats. An full overview of the processed data is reported in Table A.1 in Appendix A.

The corpus covers 17 languages, 16 belonging to the Indo-european family and one to the Uralic family (i.e., Hungarian). The length of the documents is a dimension of variation among the countries, dependent mostly on different archival practices across the countries. Some countries include only the changes to legal acts, while others combine the original text and the changes. A further peculiarity concerns the levels of detail that accompany the modalities of the measures taken, with some countries being very precise and others addressing the issue in a much broader manner. These dissimilarities can be observed by looking at the average number of sentences per country / per language. Countries such as Switzerland, Latvia, Slovenia, and Czechia appear to have the longest documents

[2] The EEA countries are: Austria, Belgium, Croatia, Cyprus, Czechia, Denmark, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Netherlands, Poland, Sweden, and Spain.

[3] All scripts, corpus, annotated data, and system(s) are available at https://github.com/tommasoc80/COVID19_emergency_event

[4] Belgium is an exception: we could collect only documents in French.

[5] https://spacy.io/universe/project/spacy-udpipe

Class Id	Class Label	# Subclasses	Example
E1	State of Emergency	18	*A restriction or requirement imposed under paragraph (1)— (a)by the Secretary of State may be varied (orally or in writing) by the Secretary of State;*
E2	Restrictions of fundamental rights and civil liberties	5	*The conditions or measures which may be specified under paragraph (2)(d) include (b)a restriction on P's activities;*
E3	Restrictions of daily liberties	10	*Where paragraph (2) applies, the Secretary of State or, as the case may be, registered public health consultant may impose on or in relation to P one or more screening requirements.*
E4	Closures / lockdown	15	*During the emergency period, no person may participate in a gathering which— (b)takes place indoors,*
E5	Suspension of international cooperation and commitments	6	*We are also working urgently to ensure international governments have sensible plans to enable the return of British and other travelers and, crucially, that they keep borders open for enough time to allow people to return home on commercial flights.*
E6	Police mobilization	14	∗*Controls will be carried out by police and municipal police.*
E7	Army mobilization	9	∗*Operation Resilience mobilizes the military and civilian personnel of all the armies, [...] who contribute to the fight against the spread of the COVID-19 epidemic in three main areas*
E8	Government oversight	6	*[The Scottish Ministers must] (a) take account of any information about the nature and number of incidents of domestic abuse occurring during the reporting period to which the review relates given to them*

Table 1: Overview of the exceptional measures' classes, including the associated number of subclasses. Examples 1–5 and 8 are extracted from UK legislative documents; examples 7 and 6, marked with an ∗, are translations from a French legislative document.

(with an average number of sentences per document ranging between 803.26 for Swiss documents in French to 525.13 for Czechia). On the other hand, Croatia, France, Italy, Norway, Hungary, Belgium, Denmark, Germany, Austria, and Sweden have the shortest documents with an average of 36.95 sentences per document, with Croatia being the shortest (3.98 sentences per document) and Ireland the longest (72.35 sentences per document). All remaining countries have lengths ranging between 129.75 sentences (Spain) and 397.16 (Cyprus).

The current corpus consists of 18,714,750 tokens. Variation in the number of tokens is quite spread, with Slovenia having more than 4 millions tokens and Norway only 6,037, followed by Germany with 12,011 tokens. Aggregating per language [6] changes the distribution of the data, leaving Norwegian as the least represented language, followed by Lithuanian with 42,761 tokens. In this setting, seven languages have more than 1 million tokens (French, Slovene, Latvian, Greek, English, Dutch, and Spanish). At this point, two remarks deserve to be made about our corpus. First, although comprehensive, our corpus is relatively small and its limited size may negatively impact the quality of subsequent analyses. We are aware of this limitation and intend to address it in future work. Second, and while the size of the documents varies per country, our corpus includes relatively short documents when compared to other types of legislation. This may be due to the specific nature of the issue at stake as COVID-19 containment measures were taken in an ad hoc, fragmented nature and were often not based on pre-existing crisis-management legislation.

4 Annotating Exceptional Measures

The identification of the exceptional measures has been conducted by applying a taxonomy of 8 classes. Note that, although the overall project focuses on a large range of subclasses, the size of the corpus and the dispersion of the subclasses in the initial phase of project presented in this paper did not allow to annotate documents at the subclass level.

Defining the taxonomy A multidisciplinary Scientific Board of 8 experts in comparative politics,

[6] Besides some inherent differences, the varieties of German used in Switzerland, Austria and Germany have been lumped together; the same has been done for French in Switzerland, France, and Belgium, English in Ireland and United Kingdom, and Italian for Italy and Switzerland.

Country	**# Docs.**	**# Sent.**	**Exceptional Classes**								**No Class**
			E1	E2	E3	E4	E5	E6	E7	E8	
Belgium	41	1,307	97	59	108	124	4	7	0	15	10,042
France	43	465	81	118	129	197	17	2	26	4	3,146
Hungary	6	95	0	1	2	3	0	1	0	0	753
Italy	72	928	66	94	126	211	1	7	1	31	6,887
Netherlands	11	171	0	0	12	58	0	0	0	0	2,153
Norway	18	277	13	6	40	58	31	0	23	5	2,040
Poland	20	95	22	6	17	34	0	1	4	5	671
UK	70	807	205	50	110	100	9	0	0	126	5,880
total	281	4,145	484	334	541	785	62	18	54	186	31,573

Table 2: Manual annotation: overview of the number of documents, sentences, and exceptional classes per country. *No Class* indicates the overall number of cases when a class is assigned the label 0.

crisis-management policies, public health, comparative law and human right law from five EU countries identified the eight classes and 83 subclasses that compose the taxonomy. The classes and their subclasses have been identified by applying both top-down and bottom-up methods. In particular, this was done by reviewing similar annotation initiatives (Cheng et al., 2020; Porcher, 2020; Hale et al., 2021) and by manually analyzing a random sample of 50 documents from the corpus. The sample contains at least one document per country. Table 1 illustrates the eight classes, the associated number of subclasses, and one example.[7] The classes cover key measures and variations in the level of coercion used in their implementation. Gathering data on implementation modalities is crucial to capture differences in policy styles that may be hidden between the use of the same term to refer to COVID-19 policy responses.

Annotating the measures A subset of 281 documents in eight languages has been selected for manual annotation. The annotation of the exceptional measures applies at sentence-level. The sample is based on the French, Polish, Dutch, English, Hungarian, Belgian, Italian, and Norwegian subcorpora. Annotators were allowed to assign as many subclasses as they consider relevant to each sentence, but with a total of eight main classes of exceptional measures. Sentences can potentially entail multiple exceptional classes, making this a multi-label annotation task. The annotation process results in eight binary annotations per sentence, with 0 if the specific class is not identified within the sentence and 1 if it is.The annotation has been conducted by three experts in political science working under the supervision of the project's Scientific Board. Since the annotators are not fluent in all languages and due to the impossibility of recruiting expert native speakers, some documents need to be translated[8] into English to be manually annotated. No inter-annotator agreement study has been conducted in this initial phase. We intend to remedy this limitation in the project's next development cycle. However, during the annotation phase, annotators met on a weekly basis to discuss ambiguous cases and the guidelines. Annotators are encouraged to propose new classes or subclasses. For a new (sub)class to be accepted, the measure should have been independently identified by the majority of the annotators. In this phase, no new classes were proposed.

Table 2 summarizes the results of the manual annotation. Differences across countries affects the distribution of the exceptional classes. Most sentences are mapped to one class only, making the absence of any measure (i.e., the 0 label) by far the most frequent. This is expected given the peculiar structure and style of legal texts. The variation in the distribution of the exceptional classes affects both the presence of specific classes in some countries and their frequency. For instance, France is the only country where all exceptional classes are present (although with varying frequencies); the Netherlands is the country with the fewest classes (only two, namely E3, indicating restrictions of daily liberties, and E4, regulating closures/lockdown); finally, Hungary is the country with the least amount of mentions of measures,

[7] The full list of the subclasses is presented in Appendix C.

[8] We used the Google Translate API.

Split	# Sent.	Exceptional Classes							
		E1	E2	E3	E4	E5	E6	E7	E8
Train	3,312	383	253	412	617	52	15	45	146
Dev	418	54	39	71	74	4	2	4	21
Test	418	47	42	62	93	6	1	5	19

Table 3: Data distributions for the train, dev and test splits.

with only seven annotations. In general, the most frequent classes are E3 and E4, while E6 and E7, involving police and army involvement respectively, are the least frequent. Although partial, the manual annotation already provides an indication of the difference and similarities of how different countries and political systems reacted to COVID-19.

5 Experiments

We have run a set of experiments to develop a tool to support the work of political scientists and other scholars to analyse the large amount of documents in the corpus. In this paper, we present the first development cycle of this tool that targets the automatic identification of the exceptional measures at class level. We opted for this setting mainly due to the low amount of positive instances and the sparseness of the subclass annotation. In particular, we investigate an array of machine learning algorithms distinguishing between feature-based (Section 5.1) and model-based (Section 5.2). The goal is to identify which approach works best for this task. Following the annotation method, the task is framed as a multi-label sentence classification task. Given the distribution of the manually annotated data per language and country, we opted to experiment directly using a multi-lingual setting. We report in Table 3 the split of the data into training, development, and test that we used for our experiments. Full details of the data distribution per country and per language is presented in Table A.2 in Appendix B

5.1 Feature-based Methods

We adopt a standard method for sentence representation, akin to the *Bag of Embeddings (BoE)* paradigm (Jin et al., 2016). In particular, we extract features from the input text to embed each sentence and use a shallow learning architecture for classifying it to (possibly more than one) related exceptional class. Representations include:

N-grams We extract *Term Frequency — Inverse Document Frequency (TF-IDF)* features based on both word and character n-grams, with $n \in \{2, ..., 5\}$ for words and $\{3, 7\}$ for characters. We deal with the sparsity of the resulting features by applying *Latent Semantic Analysis (LSA)*, by decomposing the n-gram feature matrix into its truncated singular value components (Halko et al., 2010). The overall feature extraction - decomposition pipeline transforms each sentence of our input text to a single dense vector.

Word Embeddings We utilize a multilingual version of the pretrained GloVe word vectors (Pennington et al., 2014; Ferreira et al., 2016). Word vectors from each sentence are aggregated using average pooling in order to provide a single vector representation for the sentence.

Feature vectors from both methods are concatenated and used as input to a classifier, which is trained on our supervised corpus using the sum of eight parallel binary cross-entropy losses, one per exceptional class, as to accommodate the potential existence of all classes within the same sentence. We experiment with three classifiers, namely: i) a *Support Vector Machine (SVM)* with a linear kernel, ii) a *Multi-Layered Perceptron* (MLP) with a single hidden layer and iii) a bi-directional *Gated Recurrent Unit (GRU)* neural network encoder (Chung et al., 2014) that contextualizes the input sentence before concatenating with the LSA-based feature vector and passing to the classifier. This latter method operates directly on GloVe embeddings, serving as a trainable alternative to the simple average pooling strategy used in the first two bagging approaches. For training the neural methods we use the `AdamW` optimizer (Kingma and Ba, 2017; Loshchilov and Hutter, 2017), and select hyperparameters after performing grid-search. Resulting values are: learning rate of 10^{-3}, dropout probability 0.25, weight decay of 10^{-2}, MLP hidden size of 128, GRU hidden size of 150, 100 LSA components and an early stop patience of 3 epochs. We perform model selection based on the performance in the development set, using averaged F1-score as the target metric for early stopping and report results on the test set.

5.2 Model-based Methods

In this second experiment setting, we initially followed the standard approach of fine-tuning a Transformer-based Language Model (TLM) on the annotated data. After experimenting with a range of multilingual models, we report the results of the best model, namely XLM-RoBERTa (Conneau et al., 2019). We fine-tuned XLM-RoBERTa using the `AdamW` optimizer with a learning rate of 3×10^{-5} and weight decay of 10^{-2}. Similar to the word-embedding setting, we train using a binary-cross entropy per output objective, and report results on the test set after selecting models according to their F1-score performance in the development set.

Besides obtaining state-of-the-art performances, it is known that generic TLMs suffer when applied to domains different from the one(s) used to train them. Different solutions have been proposed to address this issue, including creating new TLMs from scratch (Lee et al., 2020; Beltagy et al., 2019), using domain- or task-adaptive pretraining (Gururangan et al., 2020a; Chalkidis et al., 2020; Rietzler et al., 2020), and, more recently, developing modular domain experts (Gururangan et al., 2021). Given the peculiarity of the task and the domain of the texts, we explore the potential effectiveness of adding an intermediate training step in the performance of the language model in the downstream classification task, aiming at first adapting the pre-trained language model to the domain before fine-tuning. We thus further pre-train XML-RoBERTa using the entire collection of document composing the corpus (i.e, 6,649; Section 3). We replicate the XLM-RoBERTa pretraining process, applying the same random chances for masking and making sure that continuous spans of part-word tokens are mutually masked. We split our dataset into train-dev-test splits using $80 - 10 - 10\%$ of the documents per country (with a minimum of 2 documents for each split) and train with a masked language modeling *(MLM)* objective. The dev split has been used to select the best further trained model. We use a batch size of 16, and train for a maximum of 36 epochs, where the MLM loss saturates. Once the newly adapted model has been generated, we repeat the fine-tuning and evaluation step that we applied for the generic XLM-RoBERTa model.

6 Results

Comparative results for the supervised multi-label classification task for all methods are presented in Table 4. We evaluate against several sentence and word-level multi-label metrics and include results from a dummy baseline that always predicts negative existence of exceptional classes for all samples. We followed an evaluation approach similar to *Named Entity Recognition (NER)*, where only the positive classes are evaluated. The high accuracy metrics in the dummy case showcase the under-representation of positive classes, further highlighting the challenge of the task. However, it is apparent that all learning methods greatly aid in modeling the task, with the best performing method being the XLM-RoBERTa language model. Although marginal, the better performance of the further trained model (row *XLM(pre@36)* in Table 4) suggests the potential effectiveness of this technique.

We further evaluate the best system, *XLM(pre@36)*, for its language adaptation capacity by performing a series of zero-shot experiments. In particular, we fine-tune the XLM-RoBERTa models using the manually annotated training data from all countries/languages except one that is used for testing. With this experiment we want to identify whether our multilingual model is capable of learning cross-lingual concepts that are general enough to successfully detect the measures in new languages. This is also a strategy to check the expected performance of the trained models on the not-annotated documents of the corpus. The results of this experiment (presented in Table 5) highlight the intrinsic challenge of multilingual knowledge transfer across legal domains: even though sharing linguistic information, each country is very much bound to the specifics of their legal system. This is mostly evident in the results for Belgium and France, where even though both datasets are in French and comparable in size (see Table 2), individual scores are higher than 60% only when country specific training material is added. However, we observe that on average the zero-shot performance is on par with the feature-based baselines, hinting towards the benefits of incorporating such a multilingual zero-shot system

Model	Accuracy		Hamming Loss		F1-Score	Precision	Recall
	all(%)	events(%)	all(%)	events(%)	(%)	(%)	(%)
dummy	51.4	0.0	92.2	84.0	–	–	0.0
SVM	68.1	39.5	95.0	90.3	37.2	29.5	50.8
MLP	60.6	24.7	94.0	88.4	25.7	18.5	50.4
Bi-GRU	62.2	40.0	93.8	89.7	46.6	42.1	51.1
XLM(no-pre)	69.2	54.6	95.5	**93.5**	59.2	**62.6**	60.0
XLM(pre@36)	**71.3**	**57.7**	**95.6**	93.4	**59.8**	55.9	**62.8**

Table 4: Comparative performance of all baselines for the supervised multi-label document classification task. We report sentence and event-level (hamming) accuracy (%), as well as F1-score, precision and recall, averaged across all classes after computed independently for each of the eight binary classification tasks.

Country	F1-Score		Precision		Recall	
	zero(%)	train(%)	zero(%)	train(%)	zero(%)	train(%)
Belgium	43.7	**72.0**	55.9	**84.9**	36.6	**64.5**
Poland	58.3	**58.3**	53.3	**53.3**	66.7	**66.7**
France	31.8	**81.8**	27.0	**82.9**	39.3	**84.7**
Italy	33.5	**58.0**	43.1	**64.6**	36.9	**56.7**
Netherlands	20.6	**55.0**	37.5	**62.5**	23.6	**50.0**
Norway	15.5	**41.4**	13.5	**40.5**	18.9	**47.7**
UK	38.4	**69.0**	42.3	**69.5**	37.0	**70.4**
average	34.5	**62.2**	38.9	**65.6**	37.0	**62.9**

Table 5: F1-score, Precision, and Recall for the zero-shot (*zero*) and full fine-tuning (*train*) of the domain adapted XLM-RoBERTa model for each individual country in the manually annotated data. Hungarian documents are excluded due to their very limited size and number of positive class examples. Exceptional classes that are absent from the test split of some countries are disregarded from the calculation of the average.

in a human-in-the-loop co-annotation scenario, serving as a draft analysis that human experts can iterate over in future development cycles of the system.

7 Related Work

LegalAI has a longstanding tradition with early works dating back from the 1960s (Kort, 1957; Ulmer, 1963) and has seen the development of a variety of tasks ranging from the development of domain specific ontologies and lexica (Breukers and Hoekstra, 2004; Lame, 2005; Peters et al., 2007; Bonin et al., 2010; Francesconi et al., 2010), to automatic classification of legislative documents (Bartolini et al., 2004; Moens et al., 2007; Gonçalves and Quaresma, 2005; de Maat et al., 2010; Chalkidis et al., 2019b; Soh et al., 2019), automatic summarization (Farzindar and Lapalme, 2004; Galgani et al., 2012; Polsley et al., 2016; Feijo and Moreira, 2019), court judgment predictions (Zhong et al., 2018; Ye et al., 2018; Chalkidis et al., 2019a; Medvedeva et al., 2020), legal entities detection and classification (Cardellino et al., 2017; Leitner et al., 2020), and question answering systems (Taniguchi and Kano, 2016; Delfino et al., 2018; Kien et al., 2020).

Most of current work is embedded in the paradigm of Deep Learning, using embedding representations, neural networks or large pre-trained language models. The emergence of Deep Learning has been accompanied by a growth in specialized embedding representations for the legal domain. Similarly to other areas of applications of HLT, the legal domain has seen two waves of embedding methods. The first is has seen the application of static methods (e.g., `Word2Vec`, `GloVe`, `FastText`, `Doc2Vec`, a.o.) based on characters, words, or even documents (Ash and Chen, 2017;

Chen and Ash, 2019; Chalkidis and Kampas, 2019; Kayalvizhi et al., 2019; Noguti et al., 2020) and their integration in neural network architectures. The second wave has seen the development of contextualized representations and the generation of domain-specific transformer based language models (Chalkidis et al., 2020). Our work is related to this second wave of embedding representations, in particular, by using a generic multilingual pretrained model such as XLM-RoBERTa. On the contrary, rather than creating domain specific and multilingual language models from scratch, we applied Task Adaptive Pretraining (TAP) on the line of Gururangan et al. (2020b) to the generic XLM-RoBERTa as a strategy to boost early domain adaptation.

Similarly to previous work, we perform an Information Extraction task framed as classification problem at sentence level. Neill et al. (2017) introduces a sentence classification task aiming at extracting different modalities from financial legislative document in English. Modality plays a pivotal role in order to distinguish between what is permitted, prohibited, obliged. They are able to achieve an F1 score of .79 using a BiLSTM model and combining an ensemble of domain specific and generic word embedding representations based on `Word2Vec`. Chalkidis et al. (2018) improves along different dimensions including a more powerful Bi-LSTM system, a larger dataset, and the use of more fine-grained classes. Other works have applied the same task to sentences in German tenancy law (Waltl et al., 2017), and US and Italian regulations (Kiyavitskaya et al., 2008).

Other works have focused on the extraction of contract elements (Chalkidis et al., 2017) or text classification (Sulea et al., 2017; Wei et al., 2018; Chalkidis et al., 2019b). To the best of our knowledge, we have identified limited previous work on multilingual or cross-lingual applications to the legal domain (Galassi et al., 2020; Chalkidis et al., 2021). In this respect, in our work the multilingual dimension plays a pivotal role in the development of our approach. Given the homogeneity of the topic of the legislative documents taken into account (i.e., COVID-19 legislation), the multilingual dimension has been exploited to account for the limited amount of manually annotated documents in each language and country.

8 Conclusion and Future Work

In this paper, we have presented a new corpus and a taxonomy to identify exceptional measures implemented to counteract the COVID-19 emergency across 21 countries. A subset of the corpus (281 documents, 4,145 sentences) has been manually annotated. The data have been successfully applied to develop a fist version of a classifier based on a domain-adapted multi-lingual language model (XLM-RoBERTa) to support experts in investigation of the measures and their impact in the society. Besides the relatively small size of the training data, the final score of the system (F1 score 59.8) indicates promising applicability at the final stage of the project.

In the future, we plan to develop the project in two directions. First, we will extend the corpus to include additional countries/languages and extend the data to the regional/municipal levels. This will allow us to further adapt XLM-RoBERTa. Active learning methods can be adopted to boost the annotation process, leveraging on the fine-tuned models to auto-tag and manually review documents in batches to accelerate the annotation process. Second, we plan to develop a fine-grained version of the classifier to include the taxonomy's subclasses. Finally, we wish to further study the predictions of the classification system in the unsupervised data and identify potential cross-lingual keywords and/or topics that relate to each exceptional class separately.

Acknowledgments

This work is part of the EXCEPTIUS project, financed by Dutch Organization for Health Development and Research (ZonMw, project number: 10430032010026).

References

Elliott Ash and Daniel L Chen. 2017. Judge embeddings: Toward vector representations of legal belief. Technical report, Technical report.

Jan Auernhammer. 2020. Human-centered ai: The role of human-centered design research in the development of ai. In *Synergy - DRS International Conference 2020*, Online. Design Research Society.

Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. 2004. Automatic classification and analysis of provisions in italian legal texts: a case study. In *OTM Confederated*

International Conferences" On the Move to Meaningful Internet Systems", pages 593–604. Springer.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Christian Bjørnskov and Stefan Voigt. 2021. This time is different?—on the use of emergency measures during the corona pandemic. *European Journal of Law and Economics*, pages 1–19.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. Singling out legal knowledge from world knowledge. an nlp–based approach. In *Proceedings of the 4th Workshop on Legal Ontologies and Artificial Intelligence Techniques*, pages 39–50.

Joost Breukers and Rinke Hoekstra. 2004. Epistemology and ontology in core ontologies: Folaw and lricore, two core ontologies for law. In *Proceedings of the Workshop on Core Ontologies in Ontology Engineering (EKAW04)*.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 254–259, Valencia, Spain. Association for Computational Linguistics.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 19–28.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259, Melbourne, Australia. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019c. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer.

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

Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198.

Daniel L Chen and Elliott Ash. 2019. Case vectors: spatial representations of the law using document embeddings. *Law as Data*, 11.

Cindy Cheng, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. Covid-19 government response event dataset (coronanet v. 1.0). *Nature human behaviour*, 4(7):756–768.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Emile de Maat, Kai Krabben, and Radboud Winkels. 2010. Machine learning versus knowledge based classification of legal texts. In *Legal Knowledge and Information Systems*, pages 87–96. IOS Press.

Pedro Delfino, Bruno Cuconato, Guillherme Paulino Passos, Gerson Zaverucha, and Alexandre Rademaker. 2018. Using openwordnet-pt for question answering on legal domain. In *Proceedings of the 9th Global Wordnet Conference*, pages 105–112.

Clara Marie Egger, Raul Magni-Berton, Sebastian Roché, and Kees Aarts. 2021. I do it my way: Understanding policy variation in pandemic response across europe. *Frontiers in Political Science*, 3:17.

Sarah Engler, Palmo Brunner, Romane Loviat, Tarik Abou-Chadi, Lucas Leemann, Andreas Glaser, and Daniel Kübler. 2021. Democracy in times of the pandemic: explaining the variation of covid-19 policies across european democracies. *West European Politics*, pages 1–22.

Atefeh Farzindar and Guy Lapalme. 2004. Legal text summarization by exploration of the thematic structure and argumentative roles. In *Text Summarization Branches Out*, pages 27–34.

Diego Feijo and Viviane Moreira. 2019. Summarizing legal rulings: Comparative experiments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 313–322, Varna, Bulgaria. INCOMA Ltd.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2028, Berlin, Germany. Association for Computational Linguistics.

Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. Integrating a bottom–up and top–down methodology for building semantic resources for the multilingual legal domain. In *Semantic processing of legal texts*, pages 95–121. Springer.

Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 915–926.

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. In *Proceedings of the workshop on innovative hybrid approaches to the processing of textual data*, pages 115–123.

Teresa Gonçalves and Paulo Quaresma. 2005. Is linguistic information relevant for the classification of legal texts? In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 168–176.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2021. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020a. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020b. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, et al. 2021. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5(4):529–538.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2010. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop,*, San Diego, US. CEUR Workshop Proceedings.

Daniel Hopkins and Gary King. 2007. Extracting systematic social science meaning from text. *Manuscript available at http://gking. harvard. edu/files/words. pdf*, 20(07).

Peng Jin, Yue Zhang, Xingyuan Chen, and Yunqing Xia. 2016. Bag-of-embeddings for text classification. In *IJCAI*, volume 16, pages 2824–2830.

S Kayalvizhi, D Thenmozhi, and Chandrabose Aravindan. 2019. Legal assistance using word embeddings. In *FIRE (Working Notes)*, pages 36–39.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Nadzeya Kiyavitskaya, Nicola Zeni, Travis D Breaux, Annie I Antón, James R Cordy, Luisa Mich, and John Mylopoulos. 2008. Automating the extraction of rights and obligations for regulatory compliance. In *International Conference on Conceptual Modeling*, pages 154–168. Springer.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

Guiraude Lame. 2005. Using nlp techniques to identify legal ontology components: concepts and relations. In *Law and the Semantic Web*, pages 169–184. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Julian Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodriguez-Doncel, Artem Revenko, Sotirios Karampatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia, and Filippo Maganza. 2020. Orchestrating NLP services for the legal domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2332–2340, Marseille, France. European Language Resources Association.

James O' Neill, Paul Buitelaar, Cecile Robin, and Leona O' Brien. 2017. Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 159–168.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Mariana Y Noguti, Eduardo Vellasques, and Luiz S Oliveira. 2020. Legal document classification: An application to law area prediction of petitions to public prosecution service. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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

Wim Peters, Maria-Teresa Sagri, and Daniela Tiscornia. 2007. The structuring of legal knowledge in LOIS. *Artificial Intelligence and Law*, 15(2):117–135.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan. The COLING 2016 Organizing Committee.

Simon Porcher. 2020. Response2covid19, a dataset of governments' responses to covid-19 all around the world. *Scientific data*, 7(1):1–9.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources* *and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Doaa Samy, Jerónimo Arenas-García, and David Pérez-Fernández. 2020. Legal-ES: A set of large scale resources for Spanish legal text processing. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 32–36, Marseille, France. European Language Resources Association.

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. 2019. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Ryosuke Taniguchi and Yoshinobu Kano. 2016. Legal yes/no question answering system using case-role analysis. In *JSAI International Symposium on Artificial Intelligence*, pages 284–298. Springer.

S Sidney Ulmer. 1963. Quantitative analysis of judicial processes: Some practical and theoretical applications. *Law and Contemporary Problems*, 28(1):164–184.

Bernhard Waltl, Johannes Muhr, Ingo Glaser, Georg Bonczek, Elena Scepankova, and Florian Matthes. 2017. Classifying legal norms with active machine learning. In *JURIX*, pages 11–20.

Fusheng Wei, Han Qin, Shi Ye, and Haozhen Zhao. 2018. Empirical study of deep learning for text classification in legal document review. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3317–3320. IEEE.

Martin Wettstein. 2014. Content analysis of mediated public debates: Methodological framework for a computer-assisted quantitative content analysis. *Zurich, Switzerland: University of Zürich Institute of Mass Communication and Media Research (IPMZ)*.

Gregor Wiedemann. 2013. Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, pages 332–357.

Terry Winograd. 1997. From Computing Machinery to Interaction Design. In Peter Denning and Robert Metcalfe, editors, *Beyond Calculation: The Next Fifty Years of Computing*, pages 149–162. Springer-Verlag.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. 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 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. 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.

A Data Overview

Table 1 illustrates the basic statistics per country and per language of the processed documents in the corpus.

Country	Language	# Docs.	# Sent.	# Tokens	Vocab. Size	Avg. Sent. Length
Austria	German	240	13,041	331,924	13,577	25.45
Belgium	French	640	33,296	1,133,309	15,459	34.03
Croatia	Croatian	218	868	636,457	61,774	733.24∗
Cyprus	Greek	276	109,617	1,218,917	38,022	11.11
Czechia	Czech	43	22,581	213,113	12,303	9.43
Denmark	Danish	207	6,927	160,692	6,201	23.19
France	French	493	7,449	637,800	13,240	85.62
Germany	German	9	515	12,011	1,549	23.32
Hungary	Hungarian	150	3,430	134,906	6,965	39.33
Ireland	English	137	9,913	219,848	4,860	22.17
Italy	Italian	72	1,107	46,337	3,972	41.85
Latvia	Latvian	400	238,034	1,800,733	67,905	7.56
Lithuania	Lithuanian	30	4,579	42,761	4.187	9.33
Netherlands	Dutch	499	135,464	1,662,255	47,834	12.27
Norway	Norwegian Bokmål	18	307	6,037	1,837	19.72
Poland	Polish	274	78,274	888,000	23,150	11.34
Slovenia	Slovene	952	530,892	4,340,178	32,091	8.17
Spain	Spanish	669	86,807	1,790,097	38,168	20.62
Sweden	Swedish	220	13,801	130,014	4,920	9.42
	French	112	89,966	1,013,258	9,569	11.26
Switzerland	German	110	62,192	581,009	11,473	9.34
	Italian	112	62,397	713,278	9,660	11.43
UK	English	168	50,470	1,054,190	12,567	20.88
total	–	6,049	1561,927	18,767,124	441,283	–
average	–	263	67,909.87	815,961.91	19,186.22	52.18
median	–	207	22,581	636,457	12,303	19.72

A.1: Basic statistics of the corpus.

The average sentence length for Croatia is due to the SpaCy UDPipeline failing to correctly split sentences when end of the sentence punctuation marks are missing.

B Manual Annotation: Train, Dev, and Test Data Distribution

Country	**Split**	**# Sent.**	**Exceptional Classes**							
			E1	E2	E3	E4	E5	E6	E7	E8
Belgium	Train	1,045	81	43	78	94	4	6	0	10
	Dev	131	8	8	17	16	0	1	0	3
	Test	131	8	8	13	14	0	0	0	2
France	Train	371	65	97	108	156	14	2	21	3
	Dev	47	9	12	10	18	1	0	3	0
	Test	47	7	9	11	23	2	0	2	1
Hungary	Train	75	0	1	2	1	0	1	0	0
	Dev	10	0	0	1	0	0	0	0	0
	Test	10	0	0	0	1	0	0	0	0
Italy	Train	742	54	68	88	164	1	5	0	21
	Dev	93	8	12	23	17	0	1	1	5
	Test	93	4	14	15	30	0	1	0	5
Netherlands	Train	135	0	0	7	47	0	0	0	0
	Dev	18	0	0	1	7	0	0	0	0
	Test	18	0	0	4	4	0	0	0	0
Norway	Train	221	8	5	32	43	25	0	20	4
	Dev	28	2	0	2	8	2	0	0	1
	Test	28	3	1	6	7	4	0	3	0
Poland	Train	75	18	5	11	26	0	1	4	4
	Dev	10	3	1	3	3	0	0	0	1
	Test	10	1	0	3	5	0	0	0	0
UK	Train	648	157	34	86	86	8	0	0	104
	Dev	81	24	6	14	5	1	0	0	11
	Test	81	24	10	10	9	0	0	0	11
total	Train	3,312	383	253	412	617	52	15	45	146
	Dev	418	54	39	71	74	4	2	4	21
	Test	418	47	42	62	93	6	1	5	19

A.2: Overview of the manually annotated data.

C Full taxonomy of the classes and subclasses

Class ID	Class label	# of subclasses	Subclass labels
E1	State of Emergency	18	1. State of emergency. 2. Executive decision-making 3. Suspension of parliamentary debates 4. Suspension of elections 5. Suspension of initiatives & referendums 6. Suspension of constitutional courts 7. Suspension of legal advisory bodies 8. Suspension of ordinary courts 9. Suspension of subnational competence 10. Set up of a dedicated crisis accountability mechanism 11. Limitations to political opposition parties 12. Limitations to civil society organizations / intermediary associations 13. Extension of military powers/duties 14. Extension of police powers / duties 15. To check presence on street at any time or place 16. Powers to listen to conversations, access data of phones by police 17. Powers to enter homes to check lockdown at discretion of police 18. To check purchases in authorized shops / supermarkets
E2	Restrictions of fundamental rights and civil liberties	5	1. Restrictions of freedom of movement 2. Neighborhood lockdown 3. Restrictions of freedom of speech (including social media, excluding media) 4. Restrictions of freedom of press 5. Restrictions of freedom of association
E3	Restrictions of daily liberties	10	1. Wearing of masks 2. COVID19 tracking app 3. Self-isolation / quarantine 4. Stay at home requirements 5. Use of the self-filled form 6. Ban on private gatherings 7. Authorized radius outside home 8. Ban on visiting vulnerable groups 9. Restrictions on funerals 10. Restrictions on sport activities
E4	Closures / lockdown	15	1. Closure of venues of entertainment and culture 2. Ban on public gatherings 3. Daycare closure 4. Primary school closure 5. Secondary school closure 6. University / tertiary school closure 7. Closure of non-essential shops 8. Workspace closure 9. Restrictions on international travel 10. Restrictions on internal travel 11. Closure of bus network 12. Closure of metro / subway system 13. Closure of railway network 14. Closure of airports / international flights 15.Curfew implementation
E5	Suspension of international cooperation and commitments	6	1. Changes of asylum-seeking procedures evaluation 2. Suspension of trade agreements 3. Suspension of visa/permits delivery 4. Closure of embassies/ consulates 5. Repatriation of national citizens abroad 6. Recall of foreign troops abroad

Class ID	Class label	# of subclasses	Subclass labels
E6	Police mobilization	14	1. Federal / national force 2. Size of forces mobilized 3. Local forces 4. Size of forces mobilized 5. Transportation police 6. Size of forces deployed 7. Other additional public agents 8. Size of forces mobilized 9. Private forces 10. Size of forces deployed 11. Extension of powers, type of agents 12. Extension of power, if type 1 13. Extension of power, if type 2 14. Extension of power, if type 3
E7	Army mobilization	9	1. Support to health authority 2. Public order 3. Enforcing lockdown / curfew 4. Border protection 5. Enforcement of executive orders in civilian environment 6. Military on the street 7. Deployment of the military in public buildings 8. Deployment of the military in private buildings 9. Prison sentences for non-compliance
E8	Government oversight	6	1. Press conferences of the Executive 2. Publicity of executive measures 3. Creation of specific (ad hoc) accountability mechanism 4. Parliamentary investigation committee 5. Other investigation committee 6. Creation of certification of information by gov. system