Low-Resource Deontic Modality Classification in EU Legislation

Kristina Minkova **Shashank M Chakravarthy** **Gijs van Dijck**
Brightlands Institute for Smart Society & Maastricht Law and Tech Lab
Maastricht University, the Netherlands

Abstract

In law, it is important to distinguish between obligations, permissions, prohibitions, rights, and powers. These categories are called deontic modalities. This paper evaluates the performance of two deontic modality classification models, LEGAL-BERT and a Fusion model, in a low-resource setting. To create a generalized dataset for multi-class classification, we extracted random provisions from European Union (EU) legislation. By fine-tuning previously researched and published models, we evaluate their performance on our dataset against fusion models designed for low-resource text classification. We incorporate focal loss as an alternative for cross-entropy to tackle issues of class imbalance. The experiments indicate that the fusion model performs better for both balanced and imbalanced data with a macro F1-score of 0.61 for imbalanced data, 0.62 for balanced data, and 0.55 with focal loss for imbalanced data. When focusing on accuracy, our experiments indicate that the fusion model performs better with scores of 0.91 for imbalanced data, 0.78 for balanced data, and 0.90 for imbalanced data with focal loss.

1 Introduction

Obligations, permissions, prohibitions, rights, and/or powers are deontic modalities that commonly appear in legal documents. Precise identification of deontic modalities within legal texts holds importance for understanding and determining legal positions, responsibilities, and legal actions. It can contribute to automated compliance checking (Amor and Dimyadi, 2021; Dimyadi and Amor, 2013; Hashmi et al., 2018; Abdelmoneim, 2012), norm identification (Aires et al., 2017, 2019), and generation and analysis of legal policies and argumentation. Deontic modality classification (DMC) allows machines to efficiently interpret regulatory texts and easily adapt to changes in legislation (Abdelmoneim, 2012; Dias, 2022). Thus, by pursuing the objectives of this research, we aim to address the limitations of traditional human or rule-based approaches and manual processing by leveraging neural networks for DMC.

DMC is a Natural Language Processing (NLP) task, in which, given a text with *N* normative statements S=$(s_1, s_2, s_3, ..., s_n)$, a classifier is trained to label deontic modalities in each statement (Sun et al., 2023). Limited datasets are available for DMC. Waltl et al. (2017) constructed a dataset based on the German tenancy law containing 913 sentences in German. The Data Protection Regulation Compliance (DAPRECO) (Robaldo et al., 2020) dataset contains GDPR data represented in LegalRuleML (Athan et al., 2013). DAPRECO contains around 966 instances with 271 obligations, 76 permissions, and 619 constitutive rules. Rather than using or constructing a dataset that focuses on a particular legal topic or domain (e.g., GDPR, tenant law), we aimed to enhance the generalizability of the results by randomly extracting 200 provisions from EU regulations.

Prior DMC research that relied on machine learning is commonly based on annotated datasets with little or no information about the annotation process and the agreement scores. As far as agreement scores are reported, scores of between .74 and .82 are reported, which is higher than observed by Braun (2023), who, after examining 29 publications on legal datasets, found average agreement scores of .76 (Cohen's kappa), .675 (Fleiss' kappa), and .677 (Krippendorff's alpha), with the highest Krippendorff's alpha value not exceeding .78. Our research relies on an annotated dataset that resulted from multiple revision rounds by the annotators and an arbiter. We believe this approach yielded high-quality annotations, but with a limited number of annotated provisions. This is why we test fusion models, which are designed for low resource settings.

With this research on DMC, we aim to answer

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 149–158

the following questions -

1. Does a fusion model outperform a BERT-based model? (RQ1)

2. Which strategy (imbalanced dataset, balanced dataset, focal loss) provides the best results for the BERT-based and the fusion model? (RQ2)

To answer these questions, we explore a model of contrasting design - a pretrained language model on EUR-Lex data called LEGAL-BERT - and a fusion model that is smaller, not pre-trained, and designed specifically for low-resource text classification. The models are experimented on a manually curated multi-class sentence level classification data, with the help of three annotators, preserving different interpretations of the law and the genuine disagreement regarding the deontic class.

We report on the experiment setup, the various results, highlight the future opportunities for improving the models in order to give more insight on the challenges associated with low-resource DMC in EU legislation, and report the limitations of the proposed models.

2 Related Work

DMC has traditionally relied on rule-based approaches, resulting in a labor-intensive process prone to human biases (Wyner and Peters, 2011). Various studies have been conducted that have adopted machine-learning approaches to automatically detect deontic clauses. Sun et al. (2023) proposed a new method for BERT fine-tuning termed as DeonticBERT, classifying obligations, permissions, prohibitions, and other sentences from German tenancy law and Chinese social security policies, reported F1 scores of .89 (DeonticBERT) and .94 (BiLSTM). Liga and Palmirani (2022) found scores of .82 (obligations), .55 (permissions), .89 (constitutive rules) for the classification of obligations, permissions, and constitutive rules in 707 GDPR provisions by performing transfer learning based on BERT by leveraging the symbolic information of LegalXML formats. Bakker et al. (2022), focusing on Dutch legislation, came to .52 and .74 accuracy scores, by describing a method, by using rule-based and transformed based approach, to extract structured representations in the FLINT language. Waltl et al. (2019) and Glaser et al. (2018) evaluated a rule-based & machine learning approach, reporting an average of F1 scores of .83 when for German tenancy law, among other things, classifying duties, permissions, and prohibitions. Joshi et al. (2021) found .90 precision and recall scores, with BERT, for the classification of obligations, permissions, and prohibitions in contractual provisions. Shaghaghian et al. (2020), also focusing on contracts, achieved F1 scores of .92 for the classification of obligations by exploring various language, whereas Chalkidis et al. (2018) found F1 scores of .90 and .84 for obligations and prohibitions, respectively, in contractual provisions by utilising hierarchical BiLSTM. Dragoni et al. (2018), remarkably, reported precision scores of between .96 and 1.00 by combining various Natural Language Processing techniques towards extraction of obligations, permissions, and prohibitions in two Australian laws. Various other studies may be included that focus on classifying legal rules, including O'Neill et al. (2017), Waltl et al. (2017), Peters and Wyner (2016), Gao and Singh (2014), Francesconi (2010), de Maat et al. (2010), Francesconi and Passerini (2007), and Biagioli et al. (2005). A more complete and detailed overview will be available at van Dijck et al. (in preparation).

Our study adds to previous studies by comparing different approaches, including fusion-models for low resource settings, using focal loss to address class imbalance, performing multi-class classification based on randomly selected statutory provisions from EU regulations, and building and testing models for annotators separately due to genuine disagreement amongst the annotators.

3 Methods

The architecture and parameters of both LEGAL-BERT and the fusion model were optimized according to their performance on the data from one of the three annotators. Focal loss was used to address significant class imbalance by providing more weight to difficult-to-classify examples and less weight to examples in the majority class(es). Furthermore, the final fine-tuned versions of both models were trained separately on the three different datasets (one per annotator) for multi-class classification, where the full provisions were split into separate sentences or clauses for simplicity and where the possible labels were 'Obligation','Permission', 'Power', 'Prohibition', 'Right' or 'None' (i.e., non-annotated text).

3.1 Data Collection and Annotation

To cater to diverse domains within EU legislation in English without utilizing any translation services, we constructed a dataset by extracting 200 provisions from EU law[1]. To extract these 200 unique diverse provisions, an extractor was used by first querying the SPARQL endpoint provided by EU publications office[2] for regulations that are in force and not repealed. The random selection of articles of EU regulations was expected to enhance the generalizability of the results, at least with respect to EU legislation. On obtaining the provisions, three annotators annotated text as 'Obligation', 'Permission', 'Power', 'Prohibition', 'Right', or no annotation. The annotators were law students with different nationalities who also had received basic training in NLP. They engaged more than three months in over ten practice rounds to familiarize themselves with the annotation guidelines. Subsequently, the annotators labeled the same provisions and had the possibility to revise annotations after receiving feedback on where they disagreed. An arbiter corrected the annotations not made in accordance with the annotation guidelines. The agreement scores increased at each step. The annotation process and inter-rater and intra-rater agreement scores are reported elsewhere (van Dijck et al., in preparation). The annotation guidelines will be made available at https://doi.org/10.34894/ZJJ.

The possibility of genuine disagreement is the reason why we trained models for the three annotators separately. The first annotator, A1, made 567 annotations, 445 of which were obligations, 83 permissions, 17 prohibitions, 13 rights, and 9 power sentences. A2 annotated 561 sentences - 436 obligations, 77 permissions, 16 prohibitions, 16 rights, and 16 power sentences. A3 labeled 566 sentences - 436 obligations, 79 permissions, 16 prohibitions, 15 rights, and 20 power sentences. There were also 225 non-annotations, which were shuffled and split equally into three separate sets and added to the datasets of each annotator in order to learn the models to distinguish deontic and non-deontic sentences. Normal data distribution and neutral skewness (skewness between -0.05 and 0.05) were preferred as they generally lower the misclassification rate and bias of classifiers (Trafimow et al., 2018; Larasati et al., 2019; Liu et al., 2019). Since the datasets were highly skewed towards the 'Obligation' class (> 2 right-skewness), we undersampled (randomly) the obligations so that each dataset contained less than 100 instances of obligations, thus, reducing the skewness to around 1. The resulting balanced datasets were stored separately in order to investigate the impact of the class imbalance on the classifiers. The class distribution of each dataset is provided in Table 1.

Class	A1	A2	A3
None	75	75	75
Obligation	444 **(88)**	435 **(79)**	435 **(82)**
Permission	83	77	79
Power	9	16	20
Prohibition	17	16	16
Right	13	16	15
Skewness	2.18 **(0.99)**	2.11 **(0.95)**	2.05 **(0.93)**
Total	641 **(285)**	635 **(279)**	640 **(287)**

Table 1: Class distribution of the data per annotator (the number of instances after the sample reduction is given in parentheses)

The three datasets annotated by each of the annotators (A1, A2, A3) are referred to as A1, A2 and A3 respectively in further sections of the paper.

3.2 Classification Models

LEGAL-BERT. LEGAL-BERT (Chalkidis et al., 2020) is a variation of the BERT model (Devlin et al., 2018), having the same architecture as BERT but pre-trained from scratch on large corpora of legal texts in order to learn the domain-specific vocabulary and deeper semantics. The pre-training data includes EU and UK legislation, European Court of Justice (ECJ/CJEU) cases, European Court of Human Rights (ECtHR) cases, US court cases, and US contracts (Chalkidis et al., 2020). This in-domain knowledge of LEGAL-BERT has shown to earn significant performance gains over the fine-tuned BERT model, especially for more complex tasks such as multi-label classification (Chalkidis et al., 2020). Furthermore, Liga and Palmirani (2022) conducted experiments on deontic rule classification considering obligation, permission, constitutive and non-rule samples, which showed that, even on class-unbalanced and limited data (n = 831), the weight average accuracy of 81% produced by LEGAL-BERT outperforms a BERT

[1] https://eur-lex.europa.eu/homepage.html visited on September 20, 2023

[2] http://publications.europa.eu/webapi/rdf/sparql visited on September 20, 2023

model. We use EURLEX-BERT-BASE (uncased), trained on EUR-Lex data, for testing on our dataset.

According to Devlin et al. (2018), BERT models can be fine-tuned for downstream tasks by adding a dropout layer with a ratio of 0.1 and a classification layer of size 768. Chalkidis et al. (2020), on the other hand, suggested increasing the ratio to 0.2 for LEGAL-BERT, whereas Liga and Palmirani (2022) added two linear layers on top of the pretrained model, each activated using `ReLu` and having a dropout of 0.2, and a final output layer, activated with a `softmax` function. These three fine-tuning strategies were evaluated and compared based their performance on the data from annotator A1.

In terms of hyper-parameters, our initial experiments using the values (learning rate of 1e-5 and drop-out rate of 0.2) suggested by Chalkidis et al. (2020) indicated that there is overfitting. A batch size of 32 was unsuitable for our GPU resources and the model continued to underfit at 4 epochs. Thus, we optimized our model for learning rate $\in \{4e-6, 6e-6, 8e-6, 1e-5, 2e-5\}$, batch size $\in \{4, 8, 16\}$, and epochs $\in \{5, 6\}$. The LEGAL-BERT model was trained using Cross-Entropy (CE) loss and Adam optimizer. Given the class imbalance of the data (Lin et al., 2017), Focal loss (FL) with $\gamma \in \{0, 5\}$ with step increments of 1 was also investigated (Cao et al., 2022).

Fusion model. We investigated the fusion model architecture proposed by Maheen et al. (2022), which aims to utilize the strengths of CNN and (Bi)LSTM architectures - the spatial invariance of CNNs and the long-term dependencies capturing of RNNs, while minimizing the effect of their respective weaknesses – the limited global context and vanishing gradient. CNN layers slide a fixed-size window of length (n) over the word embeddings to extract local features, which are subsequently fed into the (Bi)LSTM layers to learn the sequential dependence of the words. The baseline model had the following structure:

CNN + Attention layer + BiLSTM + CNN

Fine-tuning efforts involved testing different fusion chain configurations by adding/removing layers and/or replacing BiLSTM with LSTM layers and searching over different layer sizes, as recommended by Maheen et al. (2022) and presented in Table 2. This was implemented by using `RandomSearch` from the `keras-tuner` library on 10 trials.

Layer	Minimal value	Maximum value	Step size
`CNN_1`	16	96	16
`CNN_1 dropout`	0.1	0.5	0.1
`Attention layer`	16	128	16
`(Bi)LSTM`	32	256	32
`(Bi)LSTM dropout`	0.1	0.5	0.1
`CNN_2`	32	256	32
`CNN_2 dropout`	0.1	0.5	0.1
`CNN_3`	32	256	32
`CNN_3 dropout`	0.1	0.5	0.1

Table 2: Layer sizes of the fusion chain

The investigated fusion variants were as follows:

1. CNN + A + BiLSTM
2. CNN + A + BiLSTM + CNN
3. CNN + A + BiLSTM + CNN + CNN
4. CNN + A + BiLSTM + CNN + BiLSTM
5. CNN + A + LSTM
6. CNN + A + LSTM + CNN
7. CNN + A + LSTM + CNN + CNN
8. CNN + A + LSTM + CNN + LSTM

where *A* stands for attention vector.

The fusion models were trained using Sparse Categorical Loss and Adam optimizer with a learning rate in the range of 2e-5 to 1e-3, following a step size of 2e-5 over 20 epochs, as suggested by Maheen et al. (2022). Preliminary tests showed that in our case the fusion performs best with smaller batch sizes (than the proposed 50), thus, we used a batch size of 4. Similar to LEGAL-BERT, we also experimented with focal loss in the fusion models. Early stopping was employed as a regularization technique with patience of 4 epochs.

4 Results

4.1 LEGAL-BERT

The best fine-tuning strategy for LEGAL-BERT was first identified. LEGAL-BERT was trained on all three available datasets (A1, A2, and A3) and its variations (imbalanced, balanced, and focal loss for imbalanced data) with the available fine-tuning strategies (namely Devlin et al. (2018)/ Chalkidis et al. (2020)/ Liga and Palmirani (2022)).

Table 3 shows that the strategy of Devlin et al. (2018) performed best across all scenarios (imbalanced, balanced, focal loss for imbalanced data) for the A1 dataset, whereas regularization of Chalkidis et al. (2020) and the additional neural layers of Liga and Palmirani (2022) had a negative impact. However, the latter two approaches benefited from using focal loss (2% increase in accuracy).

For A2 and A3, although the models yielded high accuracy, the results surpassed their respective training performance, indicating a possible prob-

Annotator	Imbalanced data	Balanced data	Focal loss
A1	**0.902**/ 0.870/ 0.850	**0.783**/ 0.675/ 0.518	**0.902**/ 0.891/ 0.876
A2*	**0.780**/ 0.759/ 0.770	0.789/ **0.800**/ 0.656	0.832/ 0.759/ **0.843**
A3*	0.865/ 0.859/ **0.891**	**0.767**/ **0.767**/ 0.522	0.818/ **0.891**/ 0.807

Table 3: Accuracy results for the datasets of different annotators according to the LEGAL-BERT fine-tuning architecture of Devlin et al. (2018)/ Chalkidis et al. (2020)/ Liga and Palmirani (2022) (highest results for each annotator are in bold); * Models trained on datasets from this annotator yield a test accuracy higher than their training accuracy

Epochs	Batch size	Lr = 4e-6	Lr = 6e-6	Lr = 8e-6	Lr = 1e-5	Lr = 2e-5
5	4	0.876	0.860	0.876	0.881	**0.902**
	8	0.876	0.876	0.876	0.870	0.850
	16	0.813	0.834	0.850	0.845	0.870
6	4	0.813	0.850	0.803	0.881	0.876
	8	0.839	0.881	0.865	0.870	0.865
	16	0.575	0.767	0.860	0.865	0.860

Table 4: Test accuracy results of the baseline LEGAL-BERT, dropout = 0.1, for different hyperparameters

lem in the training process. Even after balancing, the data had a moderate skewness of 0.9-1%, which might have caused the easily classifiable examples in the data to be allocated to the test set during the train/test split. Therefore, the performance of the LEGAL-BERT finetuned according to Devlin et al. (2018) and trained on data from A1 is considered further and evaluated.

The results of the experiments with various hyperparameters, for A1 dataset, is shown in Table 4.

Focal loss. The accuracy obtained with LEGAL-BERT could have been caused by the highly skewed dataset (i.e., a high number of instances with 'obligation' class, as seen in Table 1). To cater to this imbalanced data, the fine-tuning strategy was further investigated by replacing the 'Cross Entropy' loss function with 'Focal loss' for imbalanced A1 data. For focal loss, a gamma value of 2 was chosen. The accuracy of the model trained on the imbalanced data from A1 remained the same (0.9) and decreased with other gamma values (Table 5).

	LEGAL-BERT	FUSION
$\gamma = 0$	**0.902**	**0.896**
$\gamma = 1$	0.870	0.876
$\gamma = 2$	**0.902**	0.870
$\gamma = 3$	0.876	0.876
$\gamma = 4$	0.891	0.865
$\gamma = 5$	0.870	**0.891**

Table 5: Performance (accuracy) of LEGAL-BERT and CNN + attention + BiLSTM + CNN trained on the data from A1 with *Focal Loss* depending on the different γ values; FL is equivalent to CE for $\gamma = 0$

Class imbalance. With the A1 dataset, LEGAL-BERT performed with an average macro F1-score of 0.44 for all variations of data (imbalanced, balanced, focal loss with imbalanced data) (Table 6).

In Table 6, it can be observed that balancing the classes (manually) significantly lowered the precision of the LEGAL-BERT model (30% decrease) when recognizing the 'obligation' class. It seemingly drew the attention of the model towards the permissions and the non-deontic sentences and minimizes the false positives of these classes (around 10% higher recall than the imbalanced data). Focal loss increased the F1-score on the 'Permission' class by 6% but decreased for the 'None' class.

Furthermore, Table 6 shows that the 'Power', 'Prohibition', and 'Right' classes were completely unrecognized by the model. Even though focal loss is intended to dynamically scale the loss function in favor of minority classes, samples below 20 proved insufficient for learning the specific patterns and expressions for each label. The striking difference of 30%-40% between the macro and weighted average accuracy also indicates that the class imbalance had a considerable impact on the models' performance.

4.2 Fusion Model

Considering the architecture difference between LEGAL-BERT and fusion model, we start by identifying which fusion model architecture provided a high accuracy. According to the results in Table 7, although the performance difference between architectures was minor, the models that had the structure CNN + A + BiLSTM, with or without a second CNN layer, produced the highest accuracy -

	Precision	Recall	F1-score	Support
None	0.95/ 0.92/ 0.90	0.90/ 1.00/ 0.90	0.92/ 0.96/ 0.90	20/ 23/ 20
Obligation	0.91/ 0.61/ 0.92	0.99/ 0.94/ 0.98	0.95/ 0.74/ 0.95	141/ 18/ 141
Permission	0.80/ 0.83/ 0.82	0.80/ 0.93/ 0.90	0.80/ 0.88/ 0.86	20/ 27/20
Power	0./ 0./ 0.	0./ 0./ 0.	0./ 0./ 0.	2/ 2/2
Prohibition	0./ 0./ 0.	0./ 0./ 0.	0./ 0./ 0.	6/ 7/ 6
Right	0./ 0./ 0.	0./ 0./ 0.	0./ 0./ 0.	4/ 6/ 4
Accuracy			0.90/ 0.78/ 0.90	193/ 83/ 193
Macro avg	0.44/ 0.39/ 0.44	0.45/ 0.48/ 0.46	**0.45/ 0.43/ 0.45**	193/ 83/ 193
Weighted avg	0.85/ 0.66/ 0.85	0.90/ 0.78/ 0.90	0.87/ 0.71/ 0.87	193/ 83/ 193

Table 6: Classification report of the best-performing LEGAL-BERT models (dropout = 0.1) trained on the A1 dataset **(imbalanced/ balanced/ focal loss)**

89.6%.

Fusion architecture	Test accuracy
CNN + A + BiLSTM	**0.896**
CNN + A + BiLSTM + CNN	**0.896**
CNN + A + BiLSTM + CNN + CNN	0.886
CNN + A + BiLSTM + CNN + BiLSTM	0.876
CNN + A + LSTM	0.886
CNN + A + LSTM + CNN	0.886
CNN + A + LSTM + CNN + CNN	0.886
CNN + A + LSTM + CNN + LSTM	0.876

Table 7: Performance of different fusion architectures

Annotator	Imbalanced data	Balanced data	Focal loss
A1	0.896/ **0.904**	0.678/ **0.826**	0.880/ **0.886**
A2	0.829/ **0.896**	0.764/ **0.809**	0.808/ **0.896**
A3	**0.881**/ 0.871	0.756/ **0.814**	0.850/ **0.871**

Table 8: Evaluation results (accuracy) for the datasets of different annotators for 'CNN + A + BiLSTM' / 'CNN + A + BiLSTM + CNN' (highest results for each annotator are in bold)

The specific layer sizes and dropout ratios determined by the RandomSearch were: CNN_1 size = 48 (dropout = 0.1), attention layer size = 80, BiLSTM_1 size = 256 (dropout = 0.2), CNN_2 size = 224 (dropout = 0.3) and learning rate = 9.4e-4. These parameters were used in the subsequent analysis. Although the two fusion architectures performed similarly, extracting deeper features with an additional CNN layer proved to be beneficial, except for the imbalanced data of A3 (Table 8). There is no significant difference for the annotators, except for the fact that models trained on data from A1 overall achieved the highest accuracy results. Balancing the obligations class on average caused around a 10% decrease, whereas focal loss resulted in an approximately 1.4% decrease across datasets and model architectures.

Focal loss. Table 5 reveals that focal loss did not benefit the fusion model. However, using $\gamma = 5$, the prediction accuracy was only 0.05 lower than the benchmark. Focal loss did not make any significant difference in the overall classification performance but it did eliminate the model's responsiveness to the 'Right' label as seen in Table 9.

Class imbalance. Table 9 provides more insight into the classification metrics per label for A1. In general, the majority classes - 'None', 'Obligation', and 'Permission', were correctly classified by the models, trained on the imbalanced (with or without focal loss) and the balanced data from A1. On the other hand, due to their insufficient sample quantity, power statements were completely mislabeled.

In contrast to LEGAL-BERT, which was able to recognize only half of the labels, the fusion model performed much better achieving an average macro F1-score of 0.6 across A1 dataset variations. For smaller classes like 'Prohibition' and 'Right', the fusion model achieved 0.67 and 0.33 F1-scores respectively with the imbalanced data. The balancing of the obligations class yielded an improvement in the precision and recall of the classes 'None' and 'Prohibition' (10% increase), yet it proved detrimental for the remaining classes.

4.3 Error Analysis

The LEGAL-BERT and fusion model showed considerable overlap in the statements that were misclassified: 12/15 of the misclassified statements by the fusion model were also misclassified by the LEGAL-BERT-based model, whereas 12/19 of the misclassified statements by LEGAL-BERT overlapped with the misclassifications by the fusion model. Table 10 shows how the predicted labels differed for the imbalanced dataset.

Some misclassifications could, particularly in the fusion model, be considered genuine disagreement. For instance, the statement 'the electronic complaint form to be submitted to the odr platform shall be accessible to (...)' (predicted label

	Precision	Recall	F1-score	Support
None	0.76/ 0.91/ 0.80	0.84/ 0.95/ 0.84	0.80/ 0.93/ 0.82	19/ 22/ 19
Obligation	0.95/ 0.97/ 0.93	0.95/ 0.88/ 0.96	0.95/ 0.92/ 0.95	133/ 32/ 133
Permission	0.84/ 0.67/ 0.84	0.96/ 0.95/ 0.93	0.90/ 0.78/ 0.88	28/ 19/ 28
Power	0./ 0. / 0.	0./ 0./ 0.	0./ 0./ 0.	2/ 3/ 2
Prohibition	0.80/ 1.00/ 0.80	0.57/ 0.67/ 0.57	0.67/ 0.80/ 0.67	7/ 3/ 7
Right	0.50/ 0.50/ 0.	0.25/ 0.20/ 0.	0.33/ 0.29/ 0.	4/ 5/ 4
Accuracy			0.91/ 0.83/ 0.90	193/ 84/ 193
Macro avg	0.64/ 0.67/ 0.56	0.60/ 0.61/ 0.55	**0.61/ 0.62/ 0.55**	193/ 84/
Weighted avg	0.89/ 0.82/ 0.87	0.91/ 0.83/ 0.90	0.90/ 0.82/ 0.89	193/ 84/ 193

Table 9: Classification report of the CNN + A + BiLSTM + CNN trained on the data from annotator A1 **(imbalanced/ balanced/ focal loss)**

(fusion) = obligation, true label = right) may be considered a right (entitlement), yet it can also be interpreted as an obligation if it is read as 'shall be *made* accessible'. Sometimes, misclassifications were observed in statements with important terms in the legislative provisions that the model might not have been trained on. For instance, the combination 'only may' in 'only activities linked to the closure of the programme may be carried out between January and September' indicates an obligation (true label) whereas the predicted label by the fusion model was a permission, which would have been correct if the sentence would not have contained the word 'only'. This type of misclassification can perhaps be fixed with additional, more diverse training data. Other misclassifications did not have obvious explanations. For instance, the sentence 'list of parts or equipment which may pose a serious risk to (...) is set out in annex x to this regulation' (predicted label (fusion/LEGAL-BERT) = obligation/permission, true label = none) clearly does not contain an obligation.

Fusion	
Predicted label	*True label*
None	Obligation
None	Prohibition
Obligation	None
Obligation	None
Obligation	None
Obligation	Right
Obligation	Right
Obligation	Right
Obligation	Right
Obligation	Prohibition
Permission	Obligation
Permission	None
Permission	Power
Permission	Power
Prohibition	Permission

LEGAL-BERT	
Predicted label	*True label*
None	Prohibition
Obligation	None
Obligation	Permission
Obligation	Permission
Obligation	Permission
Obligation	Permission
Obligation	Prohibition
Obligation	Prohibition
Obligation	Prohibition
Obligation	Prohibition
Obligation	Prohibition
Obligation	Right
Obligation	Right
Obligation	Right
Obligation	Right
Permission	None
Permission	Obligation
Permission	Power
Permission	Power

Table 10: Misclassfications

5 Discussion

RQ1: Does a fusion model outperform a BERT-based model? When evaluating with macro F1-scores, it is seen that the fusion model outperforms LEGAL-BERT significantly. The similar results for the three annotators indicate that the models tested were adaptive to differences in input. Overall, the findings suggest that fusion models might be a suitable approach for testing DMC in a low-resource setting.

RQ2: Which strategy (imbalanced dataset, balanced dataset, focal loss) provides the best results for the BERT-based and the fusion model? LEGAL-BERT performed with a macro F1-score of 0.45 for imbalanced data (with obligations having higher instances), 0.43 on balanced data, and 0.45 with focal loss as loss function for imbalanced data. In comparison, the fusion model performed with a macro F1-score of 0.61 for imbalanced data, 0.62 with balanced data, and 0.55 with focal loss as loss function for imbalanced data.

As the data was skewed towards the 'Obligation' class, evaluating with macro F1-score highlights the performance of the model per available class. The fusion model performed better with an increase of 10-20% with macro F1-scores across variations of the dataset when compared to LEGAL-BERT. LEGAL-BERT saw a decrease in performance when the data was more balanced, whereas the fusion model performed consistently well across dataset variations.

It should be noted that the fusion model did not require any pre-training on large text corpora. Moreover, the fusion model exhibited a lower discrepancy between the macro and weighted average accuracy, indicating its higher resilience to class

imbalance and thus making it more suitable for low-resource settings. Solely evaluating with accuracy can misrepresent the true performance of a model.

6 Conclusion and Future Work

Previous research has tested models that were trained on large corpus or fine-tuned on datasets of a single domain with data resources ranging from 500 to 1000 instances to perform DMC. With the annotation process being time-consuming, we relied on a carefully annotated dataset of 200 legislative provisions of EU regulations that were randomly sampled and that consequently covered a variety of legal topics. Legal text can be interpreted in multiple ways, and instead of curating a single dataset based on Inter-Rater Reliability (IRR) scores, we trained multiple models for each annotator to understand the variation in model behavior.

We tested previously researched BERT-based models with low resources and evaluated them against a fusion model designed for low-resource text classification. LEGAL-BERT model and other variations through fine-tuning under performed for imbalanced data and further dropped in performance with balanced data despite the large corpora of EU Legislative text utilized for pre-training. On the other hand, a fusion model ('CNN + A + BiLSTM + CNN' variation) designed for low-resource text classification performed consistently well with both balanced and imbalanced data without the need for any pre-training on large corpora of text.

Future activities may focus on the challenging task of creating representative datasets that include more powers, prohibitions, and rights while maintaining a sufficient number of obligations and permissions. Future research may also include exploring and testing approaches that handle differences in annotations between annotators as a result of genuine disagreement (as opposed to disagreements due to mistakes). In this respect, we plan to explore the ensemble approach for models trained on each annotator to take into consideration the disagreements between annotators.

7 Ethical Implications and Limitations

The random sampling benefited the generalizability of the results yet resulted in class imbalance, as the dataset contained many obligations but few powers, prohibitions, and rights.

No ethical concerns were identified. The models were trained on publicly available data (EU legislation) and no personal information (names) of the annotators was included in any model training.

Acknowledgments. We would like to thank the anonymous reviewers for their helpful comments.

Reproducibility. The data and code will be made available at https://doi.org/10.34894/HQ8LIH, the annotation guidelines at https://doi.org/10.34894/ZJJ.

Author Contributions. Research design, SMC&KM; Data collection, GvD; Analysis, KM; writing—original draft preparation, KM&SMC&GvD; writing—review and editing, SMC&GvD&KM; supervision, SMC&GD; funding acquisition, GvD. All authors have read and agreed to the published version of the manuscript.

Funding. This research was funded by the European Union H2020 Research and Innovation Program under Grant Agreement No. 101006828—Flying Forward 2020.

References

Dareen Abdelmoneim. 2012. Semantic deontic modeling and text classification for supporting automated environmental compliance checking in construction.

João Paulo Aires, Roger Granada, Juarez Monteiro, Rodrigo Barros, and Felipe Meneguzzi. 2019. Classification of Contractual Conflicts via Learning of Semantic Representations. pages 1764–1766.

João Paulo Aires, Daniele Evelin Viana Pinheiro, Vera Lúcia Antunes De Lima, and Felipe Meneguzzi. 2017. Norm conflict identification in contracts. *Artificial Intelligence and Law*, 25(4):397–428.

Robert Amor and Johannes Dimyadi. 2021. The promise of automated compliance checking. *Developments in the built environment*, 5:100039.

Tara Athan, Harold Boley, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Wyner. 2013. Oasis legalruleml. In *proceedings of the fourteenth international conference on artificial intelligence and law*, pages 3–12.

Roos Bakker, Romy A.N. van Drie, Maaike de Boer, Robert van Doesburg, and Tom van Engers. 2022). Semantic role labelling for dutch law texts. Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 448–457. European Language Resources Association.

C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria. 2005. Automatic semantics extraction in law documents. page 133–140. Association for Computing Machinery.

Daniel Braun. 2023. I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. *Artificial Intelligence and Law*.

Lu Cao, Xinyue Liu, and Hong Shen. 2022. *Adaptable focal loss for imbalanced text classification*.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and Prohibition Extraction Using Hierarchical RNNs. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School.

Emile de Maat, Kai Krabben, and Radboud Winkels. 2010. Machine learning versus knowledge based classification of legal texts. In *Proceedings of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference*, page 87–96. IOS Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

João Dias. 2022. State of the Art in Artificial Intelligence applied to the Legal Domain.

Johannes Dimyadi and Robert Amor. 2013. Automated Building Code Compliance Checking - Where is it at? *ResearchGate*.

Mauro Dragoni, Serena Villata, Williams Rizzi, and Guido Governatori. 2018. Combining natural language processing approaches for rule extraction from legal documents. In *AI Approaches to the Complexity of Legal Systems*, pages 287–300. Springer International Publishing.

E. Francesconi and A. Passerini. 2007. Automatic classification of provisions in legislative texts. *Artificial Intelligence and Law*, 15(1):1–17.

Enrico Francesconi. 2010. Legal rules learning based on a semantic model for legislation. In *Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT-2010)*.

Xibin Gao and Munindar P. Singh. 2014. Extracting normative relationships from business contracts. page 101–108. International Foundation for Autonomous Agents and Multiagent Systems.

Ingo Glaser, Elena Scepankova, and Florian Matthes. 2018. Classifying semantic types of legal sentences: Portability of machine learning models. In *Jurix*, Frontiers in Artificial Intelligence and Applications., pages 61–70.

Mustafa Hashmi, Pompeu Casanovas, and Louis De Koker. 2018. Legal Compliance Through Design: Preliminary Results of a Literature Survey. *ResearchGate*.

Vivek Joshi, Preethu Rose Anish, and Smita Ghaisas. 2021. Domain adaptation for an automated classification of deontic modalities in software engineering contracts. page 1275–1280. Association for Computing Machinery.

Aisyah Larasati, Apif M. Hajji, and Anik Dwiastuti. 2019. The relationship between data skewness and accuracy of Artificial Neural Network predictive model. *IOP conference series*, 523(1):012070.

Davide Liga and Monica Palmirani. 2022. *Transfer Learning for Deontic Rule Classification: The Case Study of the GDPR*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Tianyu Liu, Wenhui Fan, and Cheng Wu. 2019. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial Intelligence in Medicine*, 101:101723.

Syed Mustavi Maheen, Moshiur Rahman Faisal, Md. Rafakat Rahman Karim, and Md. Shahriar. 2022. Alternative non-BERT model choices for the textual classification in low-resource languages and environments.

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. page 159–168. Association for Computing Machinery.

Wim Peters and Adam Wyner. 2016. Legal text interpretation: Identifying hohfeldian relations from text. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 379–384. European Language Resources Association (ELRA).

Livio Robaldo, Cesare Bartolini, Monica Palmirani, Arianna Rossi, Michele Martoni, and Gabriele Lenzini. 2020. Formalizing gdpr provisions in reified i/o logic: the dapreco knowledge base. *Journal of Logic, Language and Information*, 29:401–449.

S. Shaghaghian, L. Y. Feng, B. Jafarpour, and N. Pogrebnyakov. 2020. Customizing contextualized language models for legal document reviews. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2139–2148.

Jingyun Sun, Shaobin Huang, and Chi Wei. 2023. A BERT-based deontic logic learner. *Information Processing & Management*, 60(4):103374.

David Trafimow, Tonghui Wang, and Cong Wang. 2018. From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement*, 79(1):129–150.

Gijs van Dijck, Carlos Aguilera, and Shashank M Chakravarthy. in preparation. (Dis)agreement and the annotation of EU legislative provisions.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, and Florian Matthes. 2019. Semantic types of legal norms in german laws: classification and analysis using local linear explanations. *Artificial Intelligence and Law*, 27(1):43–71.

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.

Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. *International Conference on Legal Knowledge and Information Systems*, pages 113–122.