Deliverable D3.1

Report on the principles of law extraction tool

Grant Agreement No.:	101087342
Project Acronym:	POLINE
Project Title:	Principles of Law in National and European VAT
Website:	https://site.unibo.it/poline/en
Contributing WP:	[WP3]
Deliverable ID:	D3.1
Contractual delivery date:	30/09/2025
Actual delivery data:	30/09/2025
Dissemination level:	[PU]
Deliverable leader:	**[UNIBO]**

This project is funded by the European Union's Justice Programme (2022)

1 Document History

Version	Date	Author	Partner	Description
1.0	19/09/2025	Giulia Grundler	UNIBO	First draft
2.0	29/09/2025	Federico Ruggeri, Andrea Galassi, Paolo Torroni	UNIBO	Revision
3.0	30/09/2025	Giulia Grundler, Federico Ruggeri, Andrea Galassi, Paolo Torroni	UNIBO	Final Version

2 Contributors

Partner	Name	Role	Contribution
UNIBO (DISI)	Giulia Grundler, Federico Ruggeri, Andrea Galassi, Paolo Torroni	WP leaders	Development of machine learning models

3 Table of Contents

4 List of Tables

5 List of Acronyms

ML	Machine Learning
NLP	Natural Language Processing
JIF	Judicial Interpretative Formulas
VAT	Value Added Tax
TF-IDF	Term Frequency-Inverse Document Frequency
SVM	Support-Vector Machines

6 Executive Summary

This deliverable documents the Machine Learning (ML) and Natural Language Processing (NLP) models developed for the POLINE pilot tool and their results. In particular, the deliverable reports the experiments and the resulting models for achieving the WP3 task of Extraction of Principles (T3.2). This task aims at extracting principles of law in the documents produced in WP2, and connect them with related VAT concepts of the ontology. The produced models are able to process all the languages contemplated by POLINE: Italian, Bulgarian, Swedish and English.

The following table shows the data composition of the various datasets. For further information on the data and annotation process we refer to D2.2 and D2.3.

Table 1: Composition of the datasets

Language	Manually Annotated		Auto. Annotated	Total
	Test set	Validation set	Train set	
English (EU)	11	10	80	101
Bulgarian	20	20	80	120
Italian	19	19	178	216
Swedish	11	11	77	99

The developed ML models are available in, and power, the POLINE pilot tool, specifically the Customised Detection Module. This module allows recipients of VAT measures to identify judicial principles of law applied in a specific case, helping them to assess whether VAT law is correctly applied.

7 Automated extraction of JIFs

We approached the automatic extraction of JIFs using machine learning models fine-tuned on our training set. We framed the task as a binary classification: given a paragraph/sentence, classify it as a JIF or not. Experiments were conducted using the train-validation-test splits described in Table 1, determined at the document level so that paragraphs from the same document would never split across partitions. Note that the validation and test splits were entirely composed of manually annotated documents, while the training set was composed of automatically annotated data. This process ensured that the validation and test sets maintained the highest quality and avoided evaluating machine learning models using AI-annotated data.

A notable difference between English and the other languages is the unit of classification. In the English texts, the unit of classification is the paragraph, since the documents are clearly divided into numbered sections. Each paragraph is assigned a binary label (JIF or non-JIF). For the other languages, however, such a paragraph-level segmentation is not available. In these cases, we use the sentence as the unit of classification[1]: each sentence is labeled depending on whether it belongs to a possibly larger JIF or not. The task therefore remains binary in all languages, but for non-English texts the model only has access to more fine-grained, and thus more partial, information. To address this limitation, we additionally introduce a context-aware classification setting, where the model is given not only the target sentence but also its immediate context, i.e., the preceding and following sentences in the text.

Here we report, for each language, the list of models we experimented with.

English:

- DistilRoBERTa [9]: a distilled version of the RoBERTa-base model. It follows the same training procedure as DistilBERT, and it is, on average, twice as fast as RoBERTa-base.
- DeBERTa [3]: an improvement of the BERT and RoBERTa models, using disentangled attention and enhanced mask decoder.
- LEGAL-BERT [4]: a family of BERT models for the legal domain, intended to assist legal NLP research, computational law, and legal technology applications.
- LinearSVC with TF-IDF features

Bulgarian:

- SlavicBERT [1]: a version of BERT initialized on Multilingual BERT and trained on Russian News and four Wikipedias: Bulgarian, Czech, Polish, and Russian.

[1] Sentence segmentation was performed with a customized spacy pipeline that takes into account the common abbreviations of each language legal text, to avoid oversegmentation.

- BERT multilingual [5]: a pretrained model on the top 104 languages with the largest Wikipedia, using a masked language modeling objective.
- LinearSVC with TF-IDF features

Italian:

- Italian BERT [2]: an Italian version of BERT, trained on Italian Wikipedia and various texts from the OPUS corpora collection.
- ITALIAN-LEGAL-BERT [7]: a model based on Italian BERT with additional pre-training on Italian civil law corpora.
- LinearSVC with TF-IDF features

Swedish:

- Swedish BERT [8]: a BERT trained with the same hyperparameters as first published by Google, with text from various Swedish books, news, government publications, Wikipedia and internet forums.
- BERT multilingual.
- LinearSVC with TF-IDF features

For Bulgarian and Swedish a multilingual model was used due to the lack of monolingual models in these languages.

The BERT models were fine-tuned for 10 epochs with early stopping, a learning rate of 2e−5 and a batch size of 8. For reference, we also report the performance of two baselines: a classifier that outputs a random class (Random baseline) and one that always predicts the majority class (Majority baseline).

8 Results for the automated extraction of JIFs

8.1 English dataset

Table 2 reports precision, recall and F1 scores obtained by each classifier for each class, as well as their macro-average. The highest macro F1 score of 0.76 was achieved by both DistilRoBERTa and LEGAL-BERT, while DeBERTa reaches a score of 0.74. LEGAL-BERT yielded the best F1 score on the positive class (0.76) followed by DeBERTa and DistilRoBERTa. DeBERTa is the best model for what

concerns recall on the positive class, with a score of 0.82, while DistilRoBERTa reaches the maximum precision score of 0.75. For all models except DistilRoBERTa, precision is higher for the negative class while recall is higher for the positive class. LinearSVC's performance is not much inferior to state-of-the-art models, suggesting that lexical cues play a crucial role in the task.

In general, we consider LEGAL-BERT to be the best model: it has the best macro F1 score as well as the best F1 score in the positive class. Moreover, it has the second-best recall score over the positive class, close to the best. This is particularly relevant to our purpose because, in a tool intended for legal practitioners, users would want to be sure they will find what they are looking for, and if a JIF is not there, they cannot know that it is missing. In contrast, they can discard a few undesired extra paragraphs without too much effort.

Table 2: Results for the JIF classification task on the English (EU) dataset

Model	Precision			Recall			F1 score		
	yes	*no*	*avg*	*yes*	*no*	*avg*	*yes*	*no*	*avg*
Majority	0.00	0.54	0.27	0.00	1.00	0.50	0.00	0.70	0.35
Random	0.46	0.53	0.49	0.49	0.50	0.49	0.47	0.51	0.49
LinearSVC	0.68	0.76	0.72	0.75	0.70	0.72	0.71	0.73	0.72
LEGAL-BERT	0.73	0.82	0.77	0.80	0.74	0.77	**0.76**	0.77	**0.76**
DistilRoBERTa	0.75	0.79	0.77	0.75	0.78	0.77	0.75	**0.78**	**0.76**
DeBERTa	0.69	0.82	0.75	0.82	0.68	0.75	0.75	0.74	0.74

8.2 Bulgarian dataset

Table 3 reports precision, recall and F1 scores obtained by each classifier for each class, as well as their macro-average. The best macro F1 score of 0.66 is obtained by SlavicBERT in the context-aware setting. However, this setting degrades the performance of BERT multilingual from 0.64 to 0.60, showing that having access to context is not always the best option. SlavicBERT with context obtains also the best F1 score over the positive class (0.43), and the best recall over the positive class (0.46). The highest precision of 0.42 is instead obtained by BERT multilingual without context. As in the English dataset, LinearSVC exhibits not much lower performance compared to BERT models, suggesting the relevance of lexical features in this context.

We observe that classification performance is consistently lower with respect to the English language. We hypothesize that this is due to the difference in segmentation: while English texts allow paragraph-level classification, in the other languages only sentence-level units are available, which provide more limited information and likely make the task harder.

Table 3: Results for the JIF classification task on the Bulgarian dataset

Model	Precision			Recall			F1 score		
	yes	*no*	*avg*	*yes*	*no*	*avg*	*yes*	*no*	*avg*
Majority	0.00	0.85	0.43	0.00	1.00	0.50	0.00	**0.92**	0.46
Random	0.16	0.87	0.51	0.55	0.51	0.53	0.25	0.64	0.44
LinearSVC	0.42	0.87	0.65	0.19	0.96	0.57	0.26	0.91	0.59
SlavicBERT	0.39	0.89	0.64	0.38	0.90	0.64	0.38	0.90	0.64
BERT multil.	0.42	0.89	0.66	0.35	0.92	0.63	0.38	0.90	0.64
with context									
SlavicBERT	0.40	0.91	0.65	0.46	0.88	0.67	**0.43**	0.89	**0.66**
BERT multil.	0.39	0.88	0.63	0.24	0.93	0.59	0.30	0.91	0.60

8.3 Italian dataset

Table 4 reports precision, recall and F1 scores obtained by each classifier for each class, as well as their macro-average. The highest macro F1 score of 0.62 is obtained by ITALIAN-LEGAL-BERT in both context-aware and non context-aware settings. However, we consider the context-aware setting to be slightly better because it has the best F1 score on the positive class (0.32). Italian BERT exhibits slightly lower performance with respect to ITALIAN-LEGAL-BERT, higher in the setting without context.

This is the only dataset where LinearSVC achieves low performance, comparable to the majority baseline, suggesting that the Italian judgments employ a less standardized language and that lexical features alone are insufficient for accurate classification.

Moreover, as already observed for the Bulgarian dataset, classification performance is consistently lower with respect to the English language, and we attribute this mostly to the difference in segmentation.

Table 4: Results for the JIF classification task on the Italian dataset

Model	Precision			Recall			F1 score		
	yes	*no*	*avg*	*yes*	*no*	*avg*	*yes*	*no*	*avg*
Majority	0.00	0.91	0.45	0.00	1.00	0.50	0.00	**0.95**	0.48
Random	0.10	0.92	0.51	0.54	0.49	0.52	0.17	0.64	0.40
LinearSVC	0.06	0.91	0.49	0.01	0.98	0.50	0.02	0.94	0.48
Italian BERT	0.26	0.93	0.60	0.25	0.93	0.59	0.26	0.93	0.59
IT-LEGAL-BERT	0.28	0.93	0.61	0.35	0.91	0.63	0.31	0.92	**0.62**
with context									
Italian BERT	0.24	0.92	0.58	0.16	0.95	0.56	0.19	0.93	0.56
IT-LEGAL-BERT	0.29	0.94	0.61	0.38	0.91	0.64	**0.32**	0.92	**0.62**

8.4 Swedish dataset

Table 5 reports precision, recall and F1 scores obtained by each classifier for each class, as well as their macro-average. The highest macro F1 score of 0.66 is obtained by Swedish BERT in the context-aware setting. It also reaches the best f1 score on the positive class (0.37), while the highest score on the negative class belongs to the majority baseline. BERT multilingual with context is the second-best model, with a macro F1 score only 0.01 point lower than Swedish BERT (0.65). Both models without context and LinearSVC reaches comparable macro F1 scores, ranging from 0.60 to 0.62.

As previously observed with Bulgarian, the multilingual model performs worse than the language-specific one. This is a common finding in the literature, where multilingual models often underperform compared to monolingual counterparts, especially when trained on languages with limited resources or less standardized linguistic features [6,10]. Additionally, research indicates that adding large amounts of multilingual data can harm performance, likely due to limited model capacity, a phenomenon known as the 'curse of multilinguality' [11].

Moreover, as already observed for the Bulgarian and Italian datasets, classification performance is consistently lower with respect to the English language, and we attribute this mostly to the difference in segmentation.

Table 5: Results for the JIF classification task on the Swedish dataset

Model	Precision			Recall			F1 score		
	yes	*no*	*avg*	*yes*	*no*	*avg*	*yes*	*no*	*avg*
Majority	0.00	0.93	0.47	0.00	1.00	0.50	0.00	**0.97**	0.48
Random	0.08	0.95	0.52	0.62	0.51	0.57	0.15	0.66	0.41
LinearSVC	0.30	0.94	0.62	0.20	0.97	0.58	0.24	0.96	0.60
Swedish BERT	0.23	0.96	0.60	0.49	0.88	0.69	0.31	0.92	0.62
BERT multil.	0.23	0.95	0.59	0.32	0.92	0.62	0.27	0.93	0.60
with context									
Swedish BERT	0.29	0.96	0.63	0.52	0.91	0.71	**0.37**	0.94	**0.66**
BERT multil.	0.30	0.96	0.63	0.46	0.92	0.69	0.36	0.94	0.65

9 Linking JIFs to ontology concepts

The task of linking JIFs to ontology concepts starts from the extracted JIFs and aims at classifying them into one or more of the ontology concepts developed in Deliverable D2.1. It is therefore a multi-label and multi-class classification task.

Here we focus our analysis on the English dataset, since it is the only one with gold labels for this task, and therefore the only one where an evaluation of the models is possible. We focus on a subset of the ontology, particularly on the higher-level concepts: the division of the JIF concept into *non-vat* or *value added tax*, and the subelements of the latter. Of these subelements we consider the ones with at least 10 examples in the test set.

Table 6 shows the composition of our dataset. The test set is manually tagged (gold), while the validation and train set are automatically annotated with Claude 3.7. Sonnet.

Table 6: composition of the dataset of the ontology concepts used in our experiments.

ontology concept	test	val	train
non-vat	12	11	74
value added tax	90	156	996
exemptions	23	74	451
principle of fiscal neutrality	12	24	107
principle that national law must be interpreted in conformity with eu law	13	16	47
taxable amount	34	48	303
taxable persons	19	10	134
taxable transactions	13	9	192

We fine-tune LEGAL-BERT, which was the best model for the extraction task, for 10 epochs with early stopping, a learning rate of 2e−5 and a batch size of 8.

We report in Table 7 the results of the classification task, with precision, recall and F1 score of each class and their macro average. As concerns the higher-level division in *vat* and *non-vat*, the *non-vat* minority class obtains perfect precision but low recall, which leads to an F1 score of 0.50. The *vat* sub-elements reach good scores, ranging from 0.67 of *taxable transaction*, to 0.92 of *taxable persons*, with the exception of *principle that national law must be interpreted in conformity with eu law*, which obtains a score of 0.00 in both precision and recall. This low result could be attributed to the lower representation of this concept in the training set, with only 47 examples.

Despite the low result for this class, the task reaches a macro F1 score of 0.69 with 0.75 precision and 0.67 recall.

Table 7: results for the ontology concepts classification

ontology concept	precision	recall	F1 score
non-vat	1.00	0.33	0.50
value added tax	0.92	1.00	0.96
exemptions	0.78	0.91	0.84
principle of fiscal neutrality	0.75	0.75	0.75
principle that national law must be interpreted in conformity with eu law	0.00	0.00	0.00
taxable amount	0.87	0.82	0.85
taxable persons	0.94	0.89	0.92
taxable transactions	0.73	0.62	0.67
macro	**0.75**	**0.67**	**0.69**

10 Conclusion

In this deliverable, we report the ML models developed for the Extraction of Principles task in the four languages of the POLINE project. For each language, we compare a set of fine-tuned BERT models with an SVM using lexical features. We observe that classification performance in non-English languages is consistently lower compared to English. We attribute this primarily to differences in segmentation, which in turn arise from variations in text structure. Specifically, while English judgments are clearly structured and allow for paragraph-level classification, the other languages only permit sentence-level segmentation, which provides more limited information and likely increases the difficulty of the task. As for the linking of JIFs to ontology concepts, we focus our analysis on English and on the main concepts of the ontology - those sufficiently represented in the annotated dataset - achieving good results.

11 References

[1] Arkhipov M., Trofimova M., Kuratov Y., Sorokin A. (2019). Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. ACL anthology W19-3712.

[2] Bayerische Staatsbibliothek and Stefan Schweter, bert-base-italian-cased, https://huggingface.co/dbmdz/bert-base-italian-cased

[3] He, Pengcheng & Gao, Jianfeng & Chen, Weizhu. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. 10.48550/arXiv.2111.09543.

[4] 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.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[6] Kim, Y., Kim, JH., Lee, J.M. et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep 12, 13847 (2022). https://doi.org/10.1038/s41598-022-17806-8

[7] Licari, D., Comandè, G. (2024). ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain. Computer Law & Security Review, 52, Article 105908.

[8] Malmsten, Martin & Börjeson, Love & Haffenden, Chris. (2020). Playing with Words at the National Library of Sweden -- Making a Swedish BERT. 10.48550/arXiv.2007.01658.

[9] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[10] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

[11] Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.