Legal Knowledge and Information Systems
R. Markovich et al. (Eds.)

doi:10.3233/FAIA251600

Automated Extraction of Judicial Interpretative Formulas in EU Case Law on VAT

Giulia GRUNDLER [a,b,1], Piera SANTIN [c], Alessia FIDELANGELI [a], Rachele MIGNONE [a,d], Federico GALLI [a], Andrea GALASSI [b,1], Giuseppe CONTISSA [a], Luigi DI CARO [d] and Paolo TORRONI [b]

[a] *DSG, University of Bologna*
[b] *DISI, University of Bologna*
[c] *European University Institute*
[d] *Univeristy of Turin*

Abstract. This paper addresses the extraction of Judicial Interpretative Formulas (JIFs) in decisions of the Court of Justice of the European Union (CJEU) on Value Added Tax (VAT). European case law includes a significant number of JIFs on this subject, which are crucial for the interpretation of VAT. However, extracting such JIFs manually is effortful, and doing that automatically has not been investigated yet in the VAT domain. Our work proposes the first pipeline method for doing so. We start by defining a set of guidelines for annotating legal texts following a principle definition of JIF. By following such guidelines, we obtain a corpus of 21 expert-labeled CJEU decisions. We keep them for validation and testing. For training, we machine-annotate 80 additional decisions using LLMs. Our experiments show that BERT-based architectures trained on such data perform comparably to LLMs.

Keywords. Judicial Interpretative Formulas, Court of Justice of the EU, Value Added Tax, Automated Extraction, Natural Language Processing

1. Introduction

In the EU, the Court of Justice of the European Union (CJEU) is responsible for interpreting EU law and ensuring its consistent application. National courts may file requests for preliminary rulings to the CJEU, whose decisions are binding in the referring case. In fulfilling this role, the CJEU adopts an argumentative style with broad interpretative statements whose value extends beyond specific issues, as they are progressively cited and integrated into EU law [1]. Drawing on the literature [2,3], which describes *formulas* as legal texts or standards that the CJEU develops over time through self-citation, we refer to these statements as Judicial Interpretative Formulas (JIFs), highlighting their origin in case law and their interpretative function. These differ from *ratio decidendi* in that they are not limited to the elements necessary to reach the Court's conclusion [4]. One area in which the importance of JIFs is particularly evident is value-added tax. Since VAT

[1] Corresponding Authors: Giulia Grundler: giulia.grundler2@unibo.it, Andrea Galassi: a.galassi@unibo.it.

harmonisation was implemented through a Directive, regulating only certain aspects and leaving others to national discretion, numerous inconsistencies have emerged over time. Consequently, the interpretation of VAT rules has frequently fallen to the CJEU, whose rulings have progressively shaped the substance of VAT law [5].

From a technical perspective, our work builds on the field of automated extraction of legal principles, which, despite their significance, is still a largely manual process, requiring considerable effort from legal practitioners. Recent studies explore the use of automated tools to support this task, relying on hand-crafted grammars [6], machine learning [7,8,9], and prompting LLMs [10]. These approaches face two main drawbacks: data-driven methods require costly annotation, while LLM-based methods often lack stability, transparency, and reproducibility, which are essential in the legal domain.

The objective of this research is to develop and evaluate a machine-learning framework for the extraction of JIFs in CJEU case law on VAT, combining the strengths of data-driven learning with the efficiency of LLMs. We position our solution as a hybrid pipeline that aims to reduce the manual annotation burden and avoid sole reliance on a black-box LLM at deployment. First, we provide a task description that includes defining what constitutes a JIF, resulting in the development of detailed annotation guidelines. Second, we contribute a novel corpus structured into three splits and designed to separate manually annotated data for evaluation and development, and automatically annotated data for training. Third, we present an experimental study evaluating whether state-of-the-art NLP techniques can achieve acceptable results in this field. We observe that BERT models fine-tuned on LLM-annotated data are on par with LLMs in terms of performance, with the potential advantage of more stability, reproducibility, and computational efficiency. The guidelines, corpus, prompts, and code are made publicly available.[2]

2. Guidelines, Annotation Process, and Corpus

The source data consists of 101 preliminary rulings of the CJEU in the VAT domain, on the subtopics of taxable amounts and exemptions for the public interest. Of these documents, resulting from a concept-based search using the EUR-Lex Directory[3] in December 2024, 21 were manually labelled, while 80 were labelled automatically.

The annotation guidelines were drafted and iteratively improved following in-depth theoretical and empirical analysis of formulas as used in the argumentative style of the CJEU. As a general rule, we define a JIF as an interpretative statement by the Court, either formulated for the first time or drawn from cited case law. Since the objective is to extract the interpretative statements attributable to the Court, annotation focuses on the part of the decision containing the Court's answer to the preliminary question, which lies between the first occurrence of "Consideration of the question referred" (or similar expressions) and the section on litigant costs (introduced by "Costs"). The annotation unit is one paragraph, reflecting the typical structure of CJEU judgments, which are organised in numbered paragraphs, each expressing an autonomous step of reasoning [11]. In most cases, JIFs have purely legal content, i.e. interpretation of law or legal principles in abstract terms, without any reference to case-specific facts. However, JIFs may occasionally contain factual issues, such as in the final part of the decision, where

[2] https://github.com/poline-project/jif-cjeu/

[3] https://eur-lex.europa.eu/browse/directories/new-case-law.html

Table 1. Document sets manually annotated at each stage and corresponding annotation setting.

Set	Number of documents	Purpose	Annotators per document	Annotators interaction	Split of destination
A	2	guidelines draft	2	collaborative	validation
B	9	guidelines improvement	2	independent	validation
C	2	guidelines validation	2	independent	test
D	8	increase test split	1	independent	test

the Court applies its interpretation of EU law to the case at hand. These passages were also considered JIFs, as they directly express the interpretative stance of the Court. More specifically, we identify as JIFs those paragraphs containing at least one of the following: (i) interpretation of a rule, a part of a rule, or a general principle, (ii) consequences stemming from the interpretation/application of a rule or principle, (iii) subsumption of a fact within a rule, (iv) qualification of a fact as a concept contained within a rule.[4]

2.1. Manual Annotation Method

We created our guidelines and corpus through multiple stages of annotation, each involving different documents and methodologies, as reported in Table 1. The guidelines were envisioned not only as instructions for human annotators, but also as a potential source of information for automated annotation methods. To maximise their effectiveness, the guidelines included examples taken from existing documents. Since this provides a clear advantage for automatic annotation on those same texts, it required careful partitioning of the corpus to ensure sound downstream evaluation. In particular, the test split contains only documents that were never used during the development of the guidelines.

The first draft of the guidelines was based on theoretical and empirical studies. Two annotators jointly worked on a shared set of documents (set A), iteratively refining the guidelines until no ambiguities remained. To validate the guidelines, the annotators then independently annotated a new set of decisions (set B), in a double-blind process. Disagreements were discussed and resolved, leading to a consolidated version of the guidelines. A third set of decisions (set C) was then independently annotated by both annotators and used to validate this final version. We measured the Inter-Annotator's Agreement (IAA) on set C at paragraph level, restricted to the relevant portions of the judgments. The resulting Cohen's κ [12] was 0.96, indicating almost perfect agreement.

Since the guidelines were refined after annotating sets A and B, and thus relied on examples from documents contained therein, these sets were excluded from the test split and merged to form the validation split. Conversely, the guidelines remained unchanged after annotating set C, which was included in the test split, further increased by including a fourth set of decisions (set D), with each document annotated by a single expert.

2.2. Automated annotation

Given the large costs associated with manual annotation by legal experts and the need for a large annotated dataset to train BERT-based methods, we decided to automatically annotate additional documents through an LLM. We framed the task as a paragraph clas-

[4] An extended version of the guidelines, including examples, is available in our repository.

Table 2. Results of LLMs on the val split.

Model	Precision	Recall	F1
Claude	0.688	0.914	0.785
DeepSeek	0.687	0.978	**0.807**
Gemini	0.644	0.835	0.727

Table 3. Composition of the dataset.

Split	N docs	Annot.	JIF		
			Yes	No	Yes_{Avg}
Train	80	auto	1015	1037	12.7
Val	11	manual	147	164	13.4
Test	10	manual	142	164	14.2
Total	101	-	1304	1365	12.9

sification, where the model is prompted to label each paragraph as JIF or non-JIF. To reduce the computational footprint, documents were truncated to the part of the judgment where JIFs occur, as defined in the guidelines. After a preliminary study, we defined a prompt based on our guidelines and evaluated a subset of LLMs: Gemini-1.5-pro, Deepseek-R1, and Claude-3.7-sonnet.[5] Each model was compared against the manually annotated validation split to measure precision, recall, and F1 score (Table 2). DeepSeek obtained the highest score and was therefore adopted to annotate the documents.

A detailed breakdown of the final corpus composition is reported in Table 3.

3. Automated extraction

We approached the automatic extraction of JIFs as a binary classification task: given a paragraph, classify it as containing a JIF or not. Experiments were conducted using the train-validation-test splits described in Table 3, 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 manually annotated, ensuring that they maintained the highest quality and avoiding the evaluation of AI models using AI-annotated data.

We experimented with four models: **LinearSVC** with TF-IDF features, **DistilRoBERTa** [13], **DeBERTa** [14], and **LEGAL-BERT** [9]. BERT models were fine-tuned for 10 epochs with early stopping, a learning rate of $2e^{-5}$ and a batch size of 8.[6] With the exception of LinearSVC, which is deterministic, each model was trained three times with different seeds and we report their average performance. For reference, we also evaluated two baselines and **DeepSeek R1**, the generative model used to annotate the training split. The baselines consist of a classifier that outputs a random class (**Random**) and one that always predicts the majority class (**Majority**). For DeepSeek R1, we used the same setting, prompt, and hyperparameters presented in Section 2.2.

Table 4 reports the metrics obtained on the test set 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, with all other models closely behind, with a minimum score of 0.72. 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

[5] We experimented with the following hyperparameters: temperature=0.35, top-k=35, top-p=0.7.

[6] We used the following implementation of the BERT models: *distilbert/distilroberta-base*, *microsoft/deberta-v3-base*, *nlpaueb/legal-bert-base-uncased*.

Table 4. Results for the JIF classification task.

Model	Precision			Recall			F1 score			
	yes	*no*	Avg.	*yes*	*no*	Avg.	*yes*	*no*	Avg.	σ
Majority baseline	0.00	0.54	0.27	0.00	1.00	0.50	0.00	0.70	0.35	-
Random baseline	0.46	0.53	0.49	0.49	0.50	0.49	0.47	0.51	0.49	-
DeepSeek R1	0.73	0.78	0.75	0.73	0.76	0.75	0.74	0.77	0.75	-
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**	0.008
DistilRoBERTa	0.75	0.79	0.77	0.75	0.78	0.77	0.75	**0.78**	**0.76**	0.020
DeBERTa	0.69	0.82	0.75	0.82	0.68	0.75	0.75	0.74	0.74	0.012

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.

Results indicate that Transformer models generalize well, achieving performance comparable to or even better than DeepSeek. They also exhibit behaviors similar to those observed for the three generative LLMs on the validation set, namely higher recall and lower precision on the positive class, with DeBERTa showing the most pronounced difference. This can be explained considering that the models were trained over data annotated using DeepSeek. Therefore, it is reasonable to speculate that they may have learned similar patterns and a tendency to favour false positives over false negatives. Moreover, the models show stability across repeated runs, as indicated by the low standard deviation of the results. 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 is the most stable, and 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 for tools intended for legal practitioners, as the presence of additional JIFs is preferable to the absence of fundamental ones. Indeed, users can easily discard a few irrelevant paragraphs, but cannot know if a crucial JIF is missing.

4. Conclusion

In this paper, we presented a machine-learning framework for the extraction of JIFs from CJEU decisions on VAT cases. We leverage the capabilities of LLMs for annotating training data and explore the performance of various classifiers in this domain.

A general limitation of using LLMs in the legal domain lies in their proprietary nature and lack of transparency, as well as susceptibility to hallucinations [15]. However, our approach mitigates these issues by using a dedicated, task-specific classifier for the final extraction of JIFs, combining the strengths of LLM prompting with a more transparent and reproducible model. Our results indicate that the annotations produced automatically by LLMs can be exploited to distil LLM knowledge in much smaller models that obtain comparable results, reducing annotation costs and improving scalability. By enabling large-scale extraction of these statements, our approach helps various stakeholders, including judges, practitioners, and scholars, in reviewing EU case law.

Acknowledgments

This work was partially supported by the following projects: POLINE - Principles Of Law In National and European VAT (European Union's Justice Programme, G.A. No.: 101087342); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", under the European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, PE00000013.

References

[1] Jacob M. Precedents and case-based reasoning in the European Court of Justice: unfinished business. Cambridge University Press; 2014.
[2] Azoulai L. The Retained Powers' Formula in the Case Law of the European Court of Justice: EU Law as Total Law. Eur J Legal Stud. 2011;4:178.
[3] Millet FX. In the name of analogy: Judicial copy-pasting and competence creep in the connection data case law. Common Market Law Review. 2024;61(5).
[4] Komárek J. Reasoning with previous decisions: beyond the doctrine of precedent. The American Journal of Comparative Law. 2013;61(1):149-72.
[5] ARMEANIC A. Added tax: the legal practice of the European Court of Justice securing value. ACROSS Journal of Interdisciplinary Cross-border Studies. 2025 03;7.
[6] Molinari M, Quaranta M, Amantea IA. How do Case Law and Principles of Law interact, Computationally? In: Proceedings of "Workshop on Innovation and Digitization of the Justice System" - Within the XX Conference of the Italian Chapter of AIS - itAIS 2023; 2023. In Press.
[7] Shulayeva O, Siddharthan A, Wyner A. Recognizing cited facts and principles in legal judgements. Artificial Intelligence and Law. 2017;25(1):107-26.
[8] Valvoda J, Ray O. From Case Law to Ratio Decidendi. In: JURISIN. vol. 10838. Springer; 2017. p. 20-34. Available from: https://doi.org/10.1007/978-3-319-93794-6_2.
[9] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. p. 2898-904.
[10] Molinari M, Quaranta M, Amantea IA, Governatori G. Using ChatGPT to Extract Principles of Law for the Sake of Prediction: an Exploration conducted on Italian Judgments concerning LGBT(QIA+) Rights. In: Jurix '24 AI for Access to Justice Workshop; 2024. .
[11] Trklja A. The Textual Organisation of CJEU Judgments. International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique. 2024. Available from: https://doi.org/10.1007/s11196-024-10127-1.
[12] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20:37 46.
[13] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv. 2019;abs/1910.01108.
[14] He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing; 2021.
[15] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024 06;16(1):64-93. Available from: https://doi.org/10.1093/jla/laae003.