Report on the Similarity Between Principles

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:	D3.2
Contractual delivery date:	30/09/2025
Actual delivery data:	30/09/2025
Dissemination level:	PU
Deliverable leader:	**UNITO**

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

1. Document History

Version	Date	Author	Partner	Description
1.0	08/08/2025	Rachele Mignone	UNITO	Definition of the document structure
2.0	20/08/2025	Rachele Mignone, Giovanni Siragusa, Ivan Spada	UNITO	First draft
3.0	25/08/2025	Davide Audrito, Ivan Spada	UNITO	First round of revisions
4.0	10/09/2025	Rachele Mignone, Davide Audrito	UNITO	Final draft
5.0	20/09/2025	Luigi Di Caro	UNITO	Final revision

2. Contributors

Partner	Name	Role	Contribution
UNITO	Luigi Di Caro	WP Coordinator	Coordination of the Deliverable activities
UNITO	Davide Audrito	WP Participant	Contribution to the Deliverable activities
UNITO	Rachele Mignone	WP Participant	Contribution to the Deliverable activities
UNITO	Giovanni Siragusa	WP Participant	Contribution to the Deliverable activities
UNITO	Ivan Spada	WP Participant	Contribution to the Deliverable activities

3. Table of Contents

4. List of Tables

5. List of Acronyms

BERT	Bidirectional Encoder Representations from Transformers
CJEU	Court of Justice of the European Union
JIF(s)	Judicial Interpretative Formula(s)
JSON	JavaScript Object Notation
LLM(s)	Large Language Model(s)
NLP	Natural Language Processing
VAT	Value Added Tax
WP	Work package

6. Executive Summary

This deliverable establishes a scalable methodology for quantifying pairwise similarity between Judicial Interpretative Formulas. The primary contribution lies in the development and implementation of a hybrid similarity metric that synthesizes two disparate components: deep semantic similarity derived from BERT embeddings and a structural similarity based on a legal ontology. This dual approach is designed to overcome the limitations of single-paradigm metrics and to address the complexities of cross-jurisdictional legal analysis. The technical architecture, optimized for efficiency through pre-computation and the use of a symmetrical matrix, ensures that this methodology is scalable to a large legal corpus and can be deployed for real-time applications, thereby directly enabling the core functionality of the POLINE platform's analytical tools.

The deliverable is structured as follows: after the introduction (Section 7), the dataset and its creation process are described in Section 8. Section 9 describes the similarity metrics employed and their application to the JIF dataset. Finally, Section 10 describes the use of the similarity on the POLINE platform and Section 11 draws this deliverable’s conclusions.

7. Introduction

The Court of Justice of the European Union (CJEU) is responsible for ensuring the uniform interpretation and application of EU law. A notable feature of its judicial practice is a reasoning style that frequently involves the verbatim reproduction or subtle rephrasing of interpretative statements from previous judgments. This "copy-pasting"[1] or "LEGO"[2] technique has become so pronounced that the significance of a precedent often resides not in the entire judgment, but in these specific, recurring paragraphs. We refer to these passages as Judicial Interpretative Formulas (JIFs), a term introduced to identify this recognized drafting style. While prominent in EU case law, similar recurring interpretative patterns are also prevalent in national legal systems. Understanding the relationships between these JIFs is crucial for analyzing evolving trends in judicial reasoning across Europe. This Deliverable builds upon previous project tasks that first extracted and then classified JIFs, and now introduces a hybrid methodology for measuring the similarity between them to quantify their doctrinal proximity.

The POLINE project first involved automatically extracting Judicial Interpretative Formulas (JIFs) from a multilingual corpus of Value Added Tax (VAT) judgments from the CJEU and national courts in Italy, Bulgaria, and Sweden using Large Language Models (LLMs). After a systematic evaluation, Claude 3.7 Sonnet with a few-shot prompting strategy was selected, demonstrating high efficacy with an F1-score of 0.9315 on CJEU data and a precision of approximately 98.69% for Italian judgments. Following the successful extraction of thousands of JIFs, the project's second phase focused on organizing this knowledge. A multilingual VAT ontology was developed by enriching the existing EUR-LEX taxonomy with input from legal experts. LLMs were then used to classify each extracted JIF according to this ontology, transforming the flat list of formulas into a structured and conceptually organized repository of legal knowledge.

This structured and classified dataset serves as the foundation for the pairwise similarity assessment designed to measure the doctrinal proximity between any two JIFs. A simple lexical comparison is insufficient for the nuances of legal language; therefore, our methodology combines two distinct analytical dimensions.

[1] François-Xavier Millet (2024) In the name of analogy: Judicial copy-pasting and competence creep in the connection data case law.
[2] Marc Jacob (2014) Precedent application by the ECJ.

We introduce a comprehensive method for calculating the similarity between JIFs by combining two distinct dimensions. The first, semantic similarity, uses language-specific, BERT[3]-based embeddings to convert each JIF into a high-dimensional vector, capturing its contextual meaning. The second, taxonomic similarity, measures conceptual proximity by applying a taxonomy-based similarity metric to the JIFs' positions within a VAT ontology. By integrating these two scores, we create a unified, nuanced similarity metric that accounts for both the linguistic content and the formal classification of the legal statements.

8. Dataset Creation

Our analysis is based on a dataset of JIFs that were automatically extracted from legal texts using LLMs. This process was conducted on a multilingual corpus of court decisions related to VAT law, manually created by legal experts from multiple jurisdictions, including the CJEU and national supreme courts in Italy and Sweden. A systematic model selection process was implemented to identify the optimal LLM and prompt for this task, evaluating models like Claude 3.7 Sonnet[4][5], DeepSeek-R1[6][7], and Gemini 1.5 Pro[8][9] against a manually annotated dataset. This evaluation, whose results are displayed in Table 1, determined that Claude 3.7 Sonnet, combined with a few-shot prompting strategy, delivered superior performance. The extraction was framed as a binary classification task for each paragraph, where the model was prompted to return its findings in a structured JSON format to ensure machine readability. The efficacy of this method was confirmed through a quantitative evaluation by legal experts, which yielded an average F1-score of 0.932 on the CJEU test set, indicating a satisfactory ability to correctly identify JIFs. When adapted for national judgments, the methodology proved similarly effective, achieving a precision of approximately 0.987 on Italian Supreme Courts' judgments. This initial phase successfully

[3]Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[4] model snapshot: 2025-02-19. Temperature: 0.5, top-p: 0.7, top-k: 35

[5] Anthropic. (2025). *Claude 3.7 Sonnet* [Large language model]. https://www.anthropic.com/

[6] model snapshot: 2025-05-28. Temperature: 0.35, top-p: 0.7, top-k: 35

[7] DeepSeek-AI. (2025). *DeepSeek-R1* [Large language model]. https://chat.deepseek.com/

[8] model snapshot: gemini-1.5-pro-002. Temperature: 0.20, top-p: 0.7, top-k: 35

[9] Google. (2024). *Gemini 1.5 Pro* [Large language model]. https://gemini.google.com/

transformed unstructured legal texts into a reliable dataset containing thousands of extracted JIFs.

		Precision	Recall	F1-score
Zero-Shot	Gemini 1.5 pro	0.806	0.815	0.810
	Deepseek R1	0.864	0.961	**0.910**
	Claude 3.7 Sonnet	0.887	0.922	0.904
Few-Shot	Gemini 1.5 pro	0.844	0.877	0.86
	Deepseek R1	0.891	0.817	0.852
	Claude 3.7 Sonnet	0.901	0.972	**0.935**
Chain-of-Thoughts	Gemini 1.5 pro	0.829	0.76	0.793
	Deepseek R1	0.895	0.955	0.924
	Claude 3.7 Sonnet	0.874	0.972	**0.920**
Few-Shot Chain-of-Thoughts	Gemini 1.5 pro	0.865	0.707	0.778
	Deepseek R1	0.887	0.972	0.928
	Claude 3.7 Sonnet	0.884	0.983	**0.931**

Table 1. Model and prompt selection scores

The data extraction for the project yielded a corpus of 3,836 JIFs. This comprehensive dataset was composed of JIFs from three distinct jurisdictions, providing a comparative legal perspective. Specifically, the corpus included 1,402 JIFs originating from the CJEU, 478 JIFs from the Swedish jurisdiction, 445 JIFs from the Bulgarian jurisdiction, and 1,511 JIFs sourced from the Italian jurisdiction, thereby establishing a significant and heterogeneous foundation for subsequent legal analysis.

With a large corpus of JIFs extracted, the subsequent challenge was to organize this knowledge in a conceptually meaningful way to facilitate deeper analysis. To achieve this, we developed a multilingual ontology focused on the VAT domain. The initial structure of this ontology was derived from the existing EUR-LEX Directory of Case-Law taxonomy, providing an authoritative foundation. This base was then manually enriched by tax law experts who contributed additional terms and relations, expanding its breadth and depth. The resulting knowledge structure contains 127 labels (nodes) and 130 edges representing hierarchical relationships. A critical step was the creation of a multilingual dictionary to translate the ontology labels into Italian, Swedish, and Bulgarian, thereby ensuring the ontology's applicability across our entire dataset. LLMs were then employed again to classify each extracted JIF according to this new ontology. This process contextualized each formula within the broader legal domain, transforming the flat list of JIFs into a structured repository of classified legal knowledge.

9. Similarity

Following the extraction and classification of all JIFs, their pairwise similarity was assessed to provide a comprehensive understanding of their relationships within and across legal systems. This analysis was performed not only for JIFs originating from the same legal system, specifically, the European Union (Court of Justice of the European Union, CJEU), Italy, and Sweden, but also between JIFs from national judgments and those from the CJEU. To ensure the robustness of our cross-jurisdictional comparison, national JIFs were compared with their corresponding EU JIFs, which were extracted from judgments available in English. Where possible, JIFs were also obtained from the various official translations of the original EU judgment to facilitate a direct, language-specific comparison.

Our methodology for assessing similarity is built upon a two-dimensional approach, combining semantic analysis with the JIFs' taxonomic classification. This hybrid approach was designed to overcome the limitations of a purely semantic similarity evaluation, which, as we observed, could sometimes group JIFs with similar linguistic phrasing but different legal classifications. Conversely, it prevented the separation of JIFs that were conceptually and taxonomically close but used different terminology. By integrating these two dimensions, we created a comprehensive similarity metric that accurately reflects both the linguistic content and the formal legal structure of the JIFs.

9.1 Semantic Similarity

The primary dimension of our similarity assessment is based on semantic content, calculated using BERT-based embeddings that capture the contextual and semantic nuances of the legal statement. To generate these embeddings, each JIF was converted into a high-dimensional vector using a dedicated, language-specific BERT model. This approach was essential for handling the multilingual nature of our dataset.

Based on the language of the text, we used one of the following models:

- **Italian:** dlicari/Italian-Legal-BERT [10]
- **Bulgarian:** rmihaylov/bert-base-bg[11]

[10] Comande et al. *ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law,* 2022.
[11] Mihaylov, R. (2024). *rmihaylov/bert-base-bg* [Hugging Face model]. Hugging Face. Retrieved from https://huggingface.co/rmihaylov/bert-base-bg

- **Swedish:** AI-Nordics/bert-large-swedish-cased[12]
- **English:** joe32140/ModernBERT-large-msmarco[13]

Judgments from the CJEU constituted a special case. To enable accurate cross-language comparisons, the JIFs extracted from these documents were embedded in both English and all other available official translations in the projects' target languages. All linguistic versions were processed using the respective models mentioned above. This approach ensures that a robust measure of similarity can be established despite the linguistic diversity of the source documents, allowing for a direct comparison of a national JIF to its CJEU counterpart in its native language.
This semantic analysis provides a measure of linguistic similarity, effectively capturing the contextual meaning of each legal statement. However, as assessed by legal experts, a purely lexical or semantic approach was insufficient on its own, as it tended to bring JIFs with lexically similar language too close together, even when they were classified differently from a legal standpoint. This observation highlighted the need for an additional dimension of similarity.

9.2 Composite Similarity

To overcome the limitations dictated by the purely semantic approach to the similarity assessment, a new comprehensive method was implemented, employing, alongside the JIFs' content, their multilabel classification within the VAT domain.

Given a pair of JIFs (j_1, j_2) and 2 sets of classification labels $L(j_{i\in\{1,2\}}) = \{l_{ij} | \, l_{ij} \; is \; an \; ontolgy \; label \; classifying \; j_i\}$, the similarity between j_1 and j_2 can be computed as:

$$similarity\,(j_1, j_2) \;=\; cosineSimilarity(emb(j_1),\, emb(j_2)) \;*\; \frac{1}{1+\alpha * meanDistance(j_1, j_2)}$$

where

[12] AI-Nordics. (2021). AI-Nordics/bert-large-swedish-cased [Hugging Face model]. Hugging Face. Retrieved from https://huggingface.co/AI-Nordics/bert-large-swedish-cased
[13] joe32140. (2024). joe32140/ModernBERT-large-msmarco [Hugging Face model]. Hugging Face. Retrieved from https://huggingface.co/joe32140/ModernBERT-large-msmarco

- emb(j_i) is the BERT-based embedding of JIF j_i
- α is a weighting parameter, defaulting to 0.5
- the meanDistance metric is defined as the average of the shortest paths connecting all possible pairs of classification labels (l_{1i}, l_{2j}). If a non-VAT JIF is compared to a VAT JIF, their mean distance will be infinite, thus making them incomparable.

This comprehensive formula, which yields results ranging from -1 to 1, effectively integrates two distinct similarity components: the semantic and the structural. The first factor quantifies the textual similarity between the two JIFs using BERT embeddings. The second factor, structured as the inverse of a weighted distance, introduces the ontological similarity by penalizing the overall score when the JIFs' multilabel classifications are far apart within the VAT taxonomy. The adjustable weighting parameter α controls the influence of this structural difference on the final similarity score. By combining these two measures multiplicatively, the metric ensures that the assessment of legal similarity is robust, reflecting both the actual textual content and the formalized hierarchical relationship between their assigned legal concepts.

A rigorous legal assessment of the similarity results informed the final design of the POLINE platform's retrieval mechanism. To ensure the semantic relevance and practical utility of the suggested paragraphs, the minimum similarity threshold of 0.4 was defined by the legal experts' empirical assessment. This measure serves to filter out noise, guaranteeing that only JIFs with a demonstrably significant textual and conceptual overlap are displayed to the user.

Examples of the 10 most similar JIFs to Case 62014CJ0335 paragraph 32:

TEXT: “It should be observed at the outset that Article13A(1)(g) of the Sixth Directive does not specify the conditions or the procedures for recognising organisations other than those governed by public law as charitable. In consequence, it is in principle for the national law of each Member State to lay down the rules in accordance with which that recognition may be granted to such organisations (see judgment in Zimmermann, C‑174/11, EU:C:2012:716, paragraph26 and the case-law cited).”

LABELS: “organisations recognised as charitable”

Similarity	JIF text	Labels

Score		
0.9032	**sv#RÅ 2006 not 93#3** Förutom i den av Regeringsrätten åberopade domen angående Kügler har EG-domstolen i mål C-498/03 angående Kingscrest Associates Ltd m.fl. haft anledning att ta ställning till innebörden och omfattningen av artikel 13 A.1 g. Av domen framgår bl.a. följande. Syftet med det aktuella undantaget är att minska kostnaderna för tjänsterna och göra dem mer tillgängliga för de enskilda som kan få del av dem (p. 30). Enligt domstolen kan det inte anses nödvändigt att tolka uttrycket 'organisationer som är erkända som välgörenhetsorganisationer' i den aktuella artikeln särskilt restriktivt (p. 32). Vid bedömningen bör bl.a. beaktas om det föreligger särskilda bestämmelser, oavsett om dessa är nationella eller regionala, om de finns i lagar eller andra författningar, om de är förvaltningsrättsliga, skatterättsliga eller socialförsäkringsrättsliga. Vidare skall beaktas om verksamheten är av allmännyttig karaktär, om andra skattskyldiga som bedriver likadan verksamhet redan har erhållit ett liknade erkännande som välgörenhetsorganisation och om de aktuella kostnaderna eventuellt till stor del bårs av sjukkassor eller av andra socialförsäkringsorgan (p. 53). Av betydelse vid bedömningen synes också vara om aktuella organisationer underställs villkor och kontroller från de behöriga myndigheternas sida i form av registrering, inspektioner och normer avseende såväl inrättningarna som de styrande personernas kvalifikationer (jfr p. 57).	Organisations Recognised as Charitable
0.8931	**sv#RÅ 2006 not 93#2** I avgörandet RÅ 2003 ref. 21 fann Regeringsrätten att undantaget I 3 kap. 4 § första stycket ML för social omsorg inte kunde anses tillämpligt då en ekonomisk förening - i en verksamhet som saknade anknytning till offentligrättslig reglering - tillhandahöll enskilda personer samtalsterapi och själavård. I sin motivering konstaterade Regeringsrätten bl.a. att det inte fanns några indikationer på att den ekonomiska föreningens verksamhet hade en sådan anknytning till en offentligrättslig reglering som enligt EG-domstolens dom i målet C- 141/00 ang. Kügler, REG 2002 s. 1-06833, ansetts böra tillmätas betydelse. Vidare anförde Regeringsrätten att någon offentligrättslig prövning av det aktuella terapibehovet inte förekom och att de som behandlades själva bar kostnaderna	Organisations Recognised as Charitable

	för behandlingen. I domen konstaterades att den aktuella ekonomiska föreningen inte utgjorde någon erkänd välgörenhetsorganisation av det slag som avses i sjätte mervärdesskattedirektivet. Föreningen kunde inte heller med åberopande av principen om skatteneutralitet göra anspråk på skattebefrielse. Undantaget i 3 kap. .4 § första stycket ML var därför inte tillämpligt på föreningens aktuella verksamhet.	
0.8842	**sv#RÅ 2006 not 93#1** Enligt artikel 13 A.1 g i sjätte mervärdesskattedirektivet (77/388/EEG), utifrån vilken såvitt nu är aktuellt undantagsbestämmelsen för social omsorg skall tolkas, skall från skatteplikt undantas verksamhet avseende tillhandahållande av tjänster och varor som är nära kopplade till socialt bistånd eller socialförsäkring däribland sådant som tillhandahålls av ålderdomshem, offentligrättsliga organ eller andra organisationer som är erkända som välgörenhetsorganisationer av medlemsstaten i fråga.	Organisations Recognised as Charitable
0.8146	**EU#62011CJ0174#3** It is clear, however, from the case-law of the Court that, in order to determine the organisations which should be recognised as ‘charitable’ for the purposes of Article13A(1)(g) of the Sixth Directive, it is for the national authorities, in accordance with EU law and subject to review by the national courts, to take into account, in particular, the existence of specific provisions, be they national or regional, legislative or administrative, or tax or social security provisions; the public interest nature of the activities of the taxable person concerned; the fact that other taxable persons carrying on the same activities already enjoy similar recognition; and the fact that the costs of the supplies in question may be largely met by health insurance schemes or other social security bodies (see, to that effect, Kügler, paragraphs57 and58; Kingscrest Associates and Montecello, paragraph53; and, by analogy, Case C-45/01 Dornier [2003] ECR I-12911, paragraphs72 and73; L.u.P., paragraph53; and CopyGene, paragraphs65 and71).	Organisations Recognised as Charitable
0.7906	**EU#62011CJ0174#18** Moreover, applying the rules of interpretation set out in paragraph22 above (see, inter alia, Case C-473/08 Eulitz [2010] ECR I-907, paragraph42 and the case-law cited), the Court has held, in relation to the concept of ‘organisations recognised as charitable by	Organisations Recognised as Charitable

	the Member State concerned' as referred to in Article13A(1)(g) of the Sixth Directive, that that concept is in principle sufficiently broad to encompass natural persons and private profit-making entities (see Case C-216/97 Gregg [1999] ECR I-4947, paragraph17; Hoffmann, paragraph24; Kingscrest Associates and Montecello, paragraphs35 and47; and Case C-492/08 Commission v France [2010] ECR I-5471, paragraphs36 and37).	
0.6541	**it#Sez 5 Ordinanza n 22963 del 17_08_2021 (Rv 662126 01)#3** secondo la giurisprudenza comunitaria, l'art. 13, parte A, n. 1, lett. g) e h), della sesta direttiva 77/388, relativo all'esenzione dall'imposta sul valore aggiunto delle prestazioni connesse all'assistenza sociale e alla sicurezza sociale e delle prestazioni connesse alla protezione dell'infanzia e della gioventù, deve essere interpretato nel senso che la nozione di 'organismi riconosciuti come aventi carattere sociale dallo Stato membro interessato' non esclude enti privati che perseguono fini di lucro (Corte Giustizia, 26 maggio 2005, Kingscrest., C-498/03).	Welfare and Social Security Work
0.6478	**it#Sez 5 Sentenza n 11353 del 03_09_2001 (Rv 549140 01)#2** È poi irrilevante, contrariamente a quanto pure ritenuto dal giudice di appello, che la casa di riposo fosse nella specie priva delle necessarie autorizzazioni, in quanto, come riconosciuto dalla stessa Amministrazione finanziaria con la risoluzione ministeriale 28 maggio 1980 n. 382208, la norma che prevede l'esenzione in esame ha carattere oggettivo, e cioè fa riferimento per la sua applicazione al contenuto e ai destinatari delle prestazioni, indipendentemente dal previo consenso dell'ente locale competente all'esercizio dell'attività in questione.	Welfare and Social Security Work
0.6477	**it#Cass. civ., Sez. V, Sent., (data ud. 12_05_2021) 02_11_2021, n. 30975#3** in tema di Iva, ai fini dell'esenzione di cui al D.P.R. n. 633 del 1972, art. 10, 1 comma, n. 27 ter, concernente le prestazioni socio-sanitarie, di assistenza domiciliare o ambulatoriale, non è previsto il formale riconoscimento della finalità assistenziale dell'ente erogante, poichè il relativo accertamento può essere rimesso al giudice del caso concreto; nè osta all'operatività dell'esenzione la natura societaria	Welfare and Social Security Work

	dell'ente, giacchè, alla luce della giurisprudenza unionale, la nozione di 'organismi riconosciuti come aventi carattere sociale dallo stato membro' non esclude enti privati che perseguano fini di lucro (Cass. 34612/2019).	
0.6467	**it#Sez 5 Ordinanza n 22324 del 05_08_2021 (Rv 661960 01)#2** Sicchè, per quanto rileva in questa sede, le prestazioni in esame, per l'operatività dell'esenzione, devono necessariamente essere eseguite da taluno dei soggetti indicati dalla norma, a nulla rilevando, in assenza di tale presupposto soggettivo, che la prestazione sia eseguita da un differente soggetto ancorchè in forza di convenzione con uno di quelli specificati nel citato art. 10, comma 1, n. 27 ter.	Welfare and Social Security Work
0.6459	**it#Sez 5 Ordinanza n 22324 del 05_08_2021 (Rv 661960 01)#0** In tema di IVA, ai fini dell'esenzione di cui al D.P.R. n. 633 del 1972, art. 10, comma 1, n. 27 ter), norma di stretta interpretazione, necessita la realizzazione di un presupposto oggettivo, costituito dalla tipologia di prestazione in essa prevista, e di un doppio presupposto soggettivo, operante cioè tanto con riferimento al beneficiario della prestazione quanto in ordine all'esecutore della stessa (tutti rigorosamente individuati dal medesimo n. 27 ter). Ne consegue l'inapplicabilità dell'esenzione nel caso di esecuzione di prestazione, normativamente prevista, nei confronti di soggetto indicato nel citato n. 27 ter, ma da parte di soggetto diverso da taluno di quelli in esso contemplati, ancorchè in forza di convenzione intercorrente con quest'ultimo (nella specie, l'ASL).	Welfare and Social Security Work

9.3 Optimization of the Pairwise Similarity Computation

To ensure computational efficiency and manage the substantial data volume, several optimization measures were implemented for the pairwise similarity metric.

First, to prevent redundant computation, the high-dimensional embeddings for each JIF were calculated only once and persistently stored in a dedicated MongoDB collection where each JIF record is associated with an object that holds language-specific embedding arrays, allowing for efficient retrieval using the JIF's unique identifier.

Second, to adhere to this defined scope and ensure data comparability, the similarity between JIFs belonging to different jurisdictions that did not share a common language

was automatically set to zero, effectively isolating the cross-system comparisons to those with a common linguistic basis or within the EU framework.

Third, the similarity of a JIF with itself (the diagonal elements of the matrix) was set to the maximum value of one. Finally, all computed similarity scores were stored in a large symmetrical similarity matrix of size N*N (where N is the total number of JIFs) indexed by the JIFs' ids. By utilizing a symmetrical matrix, the computation only needed to be performed for half of the pairs, resulting in significant time savings. This storage strategy ensures O(1) time complexity for similarity retrieval between two existing JIFs given their IDs and maintains an efficient O(N) complexity for the insertion of a new JIF into the pre-computed structure.

9.4 Similarity Assessment in the Customized Detection Module

The Customized Detection Module (object of T4.3) is designed to operate with stringent performance requirements, necessitating rapid similarity computation for real-time user display immediately following judgment upload and JIF extraction. Due to this requisite for runtime responsiveness, the module adopts a purely semantic methodology, specifically excluding the computationally intensive ontological distance metric. This process begins with the on-demand computation of a novel BERT embedding for the uploaded JIF, utilizing a language-specific model and following the methodology described in section 9.1. Similarity is then calculated via cosine similarity between this new embedding and the entirety of the pre-computed embeddings stored in the persistent MongoDB collection (described in Section 9.3). This approach is optimized for retrieval efficiency, achieving a time complexity of O(N), where N is confined to the subset of existing JIFs that share the same jurisdiction as the document under analysis. This design prioritizes swift user feedback and the immediate utility of *ad-hoc* similarity checks.

10. Usefulness

The comprehensive pairwise similarity metric represents an advancement for jurisprudential research within the VAT domain as implemented on the POLINE platform. Its core utility lies in its direct integration with the user interface, moving beyond the limitations of traditional, keyword-based search. By immediately presenting the N most similar JIFs, this metric transforms the research process from a passive retrieval action into

a dynamic, exploratory analysis of the legal landscape.

The primary benefit for platform users is enhanced discoverability: the system ensures that relevant precedents, alternative interpretations, or comparative rulings are brought to the forefront, even if they lack common keywords with the initial search term. This functionality allows legal scholars and practitioners to rapidly juxtapose highly similar JIFs, facilitating the analysis of subtle differences in judicial reasoning, the evolution of specific legal concepts, and the identification of potentially contradictory applications of the same statute. Ultimately, the metric transforms the platform into a sophisticated analytical engine that not only retrieves information but also actively aids in the comparative analysis essential for high-quality jurisprudential research.

11. Conclusion

This deliverable has presented a robust and scalable methodology for computing pairwise similarity between Judgement Forms within the POLINE project. By leveraging a hybrid approach that combines the deep semantic understanding of BERT embeddings with the structured knowledge of a legal ontology, we have engineered a metric that is both semantically rich and contextually grounded. The technical architecture, which includes the one-time computation of embeddings stored in a persistent database and the use of a symmetrical matrix, ensures high computational efficiency, achieving near real-time performance essential for the Customized Detection Module. The application of a final similarity threshold ensures the platform consistently delivers a set of jurisprudentially relevant documents to the user. This work not only provides a foundational technical component for the POLINE platform but also demonstrates a transferable methodology for applying advanced computational linguistics and knowledge representation to complex legal datasets. The approach represents a significant step towards creating intelligent legal systems that actively assist in the discovery and comparative analysis of legal scholarship.