Gender Violence in Numbers: Prompting Italian LLMs to Characterize Crimes Against Women

Giulia Rizzi[1], Daniel Scalena[1,2] and Elisabetta Fersini[1,*]

[1]University of Milano-Bicocca, Milan, Italy
[2]University of Groningen, CLCG, Groningen, The Netherlands

Abstract

This paper investigates the application of various prompting strategies and Italian-language large language models (LLMs) to extract salient characteristics of gender-based crimes from judicial courtroom decisions. Recognizing the complex linguistic and legal structures inherent in such documents, we evaluate several types of prompting across multiple LLMs fine-tuned or pretrained on Italian corpora. Our approach focuses on identifying key elements such as crime typology, victim-perpetrator relationships, modus operandi, and main motivations behind the crimes against women. We present a comparative analysis of LLM performance on a small set of judicial courtrooms, highlighting the impact of prompt design on the extraction of legally and socially relevant information. The findings demonstrate the potential of prompt engineering to enhance the ability of LLMs to support socio-legal research and policy development in the context of gender-based violence.

Keywords

Gender violence, Information extraction, Italian court rulings, Language Models, CLiC-it

1. Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in a variety of natural language processing (NLP) tasks, showing potential for transforming domains that rely heavily on unstructured textual data [1]. In this field, the legal sector is distinguished by its unique challenges and opportunities, which can be attributed to the complexity, formalism, and high-stakes nature of judicial language.

Despite their general proficiency, LLMs remain largely untested in such highly specialized applications where linguistic nuances and factual accuracy are paramount. The extraction of structured information from legal documents, such as the personal information of the accused, necessitates not only an advanced understanding of the language, but also strict adherence to domain-specific taxonomies and ethical considerations regarding data sensitivity. The anonymised and variable structure of legal texts further complicates this task, necessitating the development of tailored strategies for effective model deployment. Beyond their technical relevance, such advancements are of considerable societal value given their potential to underpin large-scale analyses of sociological and criminological trends.

This work investigates the use of LLMs to automate the extraction of key information from anonymised court rulings in the Italian judicial system. The study's primary objectives are firstly to explore the role of prompt engineering in guiding the model's behaviour and improving output fidelity and secondly to evaluate the feasibility of using these extracted outputs to generate statistical analyses of juridical court rulings. A thorough evaluation of multiple models and prompt strategies has been undertaken, enabling the identification of both the capabilities and limitations of state-of-the-art LLMs in the context of complex, structured information retrieval within the legal domain.

The contributions of this study can be summarised as follows:

- **Prompt Evaluation** – We performed a systematic evaluation and selection of prompts tailored to a legal taxonomy, identifying the linguistic and semantic limitations that affect model performance.
- **Empirical Assessment of LLM Outputs** – We perform a detailed analysis of model behavior across multiple dimensions of a legal information extraction task, highlighting typical failure modes and model biases.
- **Data-Driven Legal Insights** – We uncover statistical trends in italian criminal justice, while emphasizing the importance of post-extraction validation due to the inherent risks of misinterpretation or hallucination, especially on such anonymised data.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
g.rizzi10@campus.unimib.it (G. Rizzi);
elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)

2. Related Works

Information extraction Information Extraction (IE) is a foundational task in natural language processing that aims to automatically extract structured information such as named entities, events, and relationships from unstructured text. Traditional IE pipelines often rely on rules or shallow machine learning models [2, 3], but recent advances have significantly improved the field, introducing more sophisticated training procedures and complex pipelines that leverage models' embedding capabilities [4, 5]. With the advent of large language models, especially generative ones, there is a growing shift toward end-to-end approaches that require minimal task-specific supervision.

In legal domain The legal domain presents unique challenges for information extraction due to its specialized terminology, complex document structures, and domain-specific entity types and relationships [6, 7, 8]. Recent studies have examined the potential of LLMs for legal IE tasks [9, 10]. These works highlight the difficulty of identifying entities such as case participants, legal concepts, and procedural events due to the prevalence of cross-references, frequent amendments, and highly specialized jargon [11, 12].

Legal documents from different jurisdictions or legal systems introduce further complications, as they may follow distinct conventions, terminologies, and structural norms, making domain transfer particularly challenging [13]. Most current language models are primarily trained on English-language data, largely sourced from Western, English-speaking jurisdictions (e.g., the United States and the United Kingdom). Research has shown that LLM performance on legal IE tasks can vary significantly between in-domain and out-of-domain contexts, with performance degradation often linked to differences in document formality, legal drafting templates, and jurisdiction-specific clauses [14]. The intricate nature of legal texts adds another layer of complexity, as legal terminology and document structures can vary widely across legal systems and languages, necessitating specialized methods for handling non-English legal texts.

Most existing work has focused on English legal documents. To the best of our knowledge, while some attempts have been made in the Italian legal domain [15, 16], no prior work has specifically addressed Italian court rulings, whose structure and terminology differ significantly from those of the Anglo-Saxon legal tradition.

3. Method

In this section we describe the introduced pipeline to extract information out of italian criminal court rulings.

3.1. Model selection

Modern language models are typically trained on vast amounts of data to capture various linguistic patterns. However, especially in the case of smaller models, the training data is often heavily skewed toward English, resulting in reduced performance on other languages. As discussed in Section 2, relatively few studies have investigated the intersection of non-English languages and legal domains. For this reason, we began by selecting models whose pre-training process or fine-tuning includes at least some Italian-language data, so as to guarantee a minimal level of competence in Italian. In particular, we evaluated three instruction-tuned checkpoints: (i) LLaMA 3.1 8B[1] [17]; (ii) Anita[2] [18], a further Italian-specific fine-tune of LLaMA 3.1 8B; and (iii) Phi-3-mini (4B parameters), instruction-tuned variant. All three models were probed on a representative subset of prompts designed to test instruction-following and the ability to emit precisely structured text suitable for information-extraction. Despite being the smallest model and having predominantly English training data, Phi-3-mini consistently produced the best-structured italian outputs and therefore emerged as the top performer in this preliminary screening.

3.2. Prompts

A campaign was designed to study several prompt engineering techniques to optimise the model's responses to the extraction task. The following prompts types have been investigated:

1. **Direct Instruction Prompt**: This type of prompt directly asks for specific information or task completion, with clear, unambiguous instructions. It's straightforward and expects a precise answer. For example: "*What is the victim's name?*".
2. **Socratic Prompt**: This type of prompt encourages Socratic reasoning by asking consequent questions. The goal is to guide the model toward discovering information or coming to conclusions. For example: "*What is the victim's name?*" followed by "*What is <name>'s gender?*".
3. **Structured Prompt**: This type of prompt provides a specific framework or format in which the response should be structured. The adopted JSON-like format includes predefined fields into which the information should be extracted. This ensures consistency and organization in the answers. For example: "*Extract the following details: {victim_name: ?, victim_gender: ?}*".

[1] `meta-llama/Llama-3.1-8B-Instruct`
[2] `swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA`

According to the selected types, 145 prompts have been defined, both manually and by utilising Large Language Models (LLMs)[3].

3.3. Dataset

To construct a suitable dataset for our study, 2,000 anonymized judicial court rulings were extracted from the DeJure corpus[4] based on the presence of references to specific norms related to gender-based crimes, i.e. Art. 609-quinquies, art. 572, art. 582, art. 609-bis, art. 609-octis, art. 609-ter, art. 612-bis of the Italian Penal Code. We engaged 5 judicial experts to finally select only those judicial court rulings effectively relevant for the considered case study. This targeted extraction strategy was employed to ensure the relevance of the selected court rulings to the legal domain under investigation. From this initial pool, a subset of 1,000 court rulings was subjected to manual evaluation by legal domain experts. The experts assessed each sentence for its appropriateness and relevance, ultimately identifying 865 court rulings as suitable for inclusion in the final dataset. This process ensured both the domain specificity and the quality of the data used in subsequent analyses. The dataset obtained has been used for the identification of pertinent information and for the extraction of statistics to finally model the gender-base violence phenomenon.

Furthermore, in order to assess the ability of the selected models to extract salient information from the court rulings, we created a subset of de-anonymisation judicial court rulings. This process was aimed at reconstructing the removed/obscured information - such as proper names, places, entities or other identifying references - by relying exclusively on the available textual content. The de-anonymization process was aimed at creating a small benchmark for qualitative analysis to compare the performance of the Italian large language models. Specifically, the original anonymised court rulings have been annotated to introduce pseudo-real information that the models could extract, in order to simulate a plausible context of application of the model itself. The de-anonymised court rulings are utilised to evaluate the capabilities of the selected models, as well as to identify the most effective prompts for the task of extracting the information included in the taxonomy.

De-anonymisation A subset of anonymized court rulings was initially subjected to a de-anonymization process using the considered language models. Each model was prompted to infer the missing information, such as names of individuals, organizations, locations, and other identifying details, based on the surrounding textual context, with the goal of filling in the fields marked as "*OMISSIS*". However, an analysis of the model outputs revealed an overall unsatisfactory quality of de-anonymization. While the models demonstrated certain inferential capabilities, the generated outputs frequently proved to be inaccurate, incomplete, or contextually inconsistent. The most critical issues arose in the reconstruction of personal names: models frequently suggested names that were inconsistent with the grammatical gender used in the text, leading to uncoherent court rulings. For instance, masculine names have been observed to be used in instances where feminine pronouns or adjectives were employed, thereby compromising the document's natural flow and readability. Furthermore, the models demonstrated inconsistency in the attribution of names throughout the document, frequently assigning different names to the same individual across multiple mentions. The absence of global coherence indicated a restricted contextual awareness, thereby diminishing the dependability of the automated procedure. In light of the aforementioned limitations, manual de-anonymization was ultimately deemed the preferred approach in order to ensure both accuracy and internal consistency.

The manual de-anonymisation process enabled the introduction of specific cases, designed to provide a thorough and robust evaluation of the models.

Foreign names were introduced to assess the models' ability to handle information that deviates from conventional paradigms. The incorporation of such cases into the study was intended to assess the models' capacity to process unconventional information and to ensure consistency and accuracy, even in the presence of elements that fall outside the more prevalent data categories utilised during their training.

Additionally, complex cases involving multiple individuals sharing the same surname were included to assess the models' ability to disambiguate identities, especially in cases where roles differ, such as a victim and defendant with the same surname. This required the models to correctly infer identities based on contextual details. Lastly, a case without any personal data was included with the objective of evaluating the efficacy of the selected models in discerning instances wherein the requested data is notably absent. The inclusion of this particular type of input allows to assess the models' ability to handle situations in which information is either completely missing or deliberately omitted.

The de-anonymisation procedure, enriched by these particular cases, results in a small dataset of 10 judicial courtroom decisions that is well-suited for the evaluation of the models' performance in challenging and incomplete scenarios.

The first dataset (composed of 865 anonymized judicial

[3] Manually generated prompts have been included as examples in the definition of a few-shot instruction to ask Chat-GPT to generate new ones.

[4] www.dejure.it

court rulings) was used to extract statistical insights on gender-based violence in Italian court rulings, while the second one composed of 10 de-anonymized court rulings served to evaluate the models' ability in the task of automatic information extraction and for the selection of the most promising prompts to adopt for the extraction task.

The understanding of crimes against woman starting from judicial courtroom decisions presents significant challenges, primarily due to the inherent complexity of legal language, which often involves dense, formal phrasing and domain-specific terminology. Additionally, judicial court rulings typically span between 3 to 15 pages (averaging about 21,000 characters, with the longest surpassing 137,000), resulting in lengthy and unstructured documents that demand robust document-level understanding. Compounding the difficulty is the frequent occurrence of multiple crimes described across different temporal contexts within a single sentence, requiring fine-grained temporal reasoning and event disentanglement to accurately identify and extract relevant legal information.

3.4. Taxonomy

A taxonomy has been defined in order to model all the relationships that are useful for the definition of the offence and the relevant entities. The objective is to obtain a complete and valid characterisation of the analysed court rulings. In order to achieve the desired taxonomy, the various classifications defined and proposed by the *Istituto Nazionale di Statistica* (ISTAT) were adopted and subsequently grouped into categories. Additional information about the identified categories, along with a schematic representation, are reported in Appendix A.

The proposed taxonomy has been adopted in the definition of the prompts for the extraction of salient characteristics of gender-based violence.

3.5. Inference pipeline description

We prompt the selected models to extract relevant information from court rulings. To ensure reproducibility, we use greedy decoding and, apply the model's original chat template from its instructed version.

A key challenge in prompting models with court rulings is their length in tokens, which can significantly slow down the generation process. Since we query the same model multiple times on the same ruling using different prompts, we leverage the decode-only nature of language models by precomputing the key-value cache for each token in the ruling. At inference time, this allows us to avoid redundant computation of internal states during each forward pass.

Each prompt includes a predefined set of labels from which the model is expected to choose based on the extracted information. The model should output at least one label, optionally accompanied by an explanation or the relevant text span. For evaluation, we perform an exact string match between the stripped model output and the set of possible labels.

4. Discussion

The selected models has been evaluated on the de-anonymized subset of court rulings focusing both on model performances and computational requirements.

Furthermore, results analysis allowed for the selection of the most promising prompts.

4.1. Prompts Evaluation

The selection of prompts played a pivotal role in determining the effectiveness of the selected language models in extracting structured information from juridical court rulings. This phase of experimentation revealed not only the variability in the interpretative capabilities of large language models (LLMs), but also several intrinsic limitations related to prompt design and the models' generalization ability when confronted with legal language.

Preliminary analyses were conducted on the manually de-anonymised subset of court rulings, which permitted the empirical identification of prompt configurations that were optimally suited to the information extraction task. This experiment was able to shed light on a number of difficulties encountered by the models. In many cases, LLMs exhibited a fundamental misunderstanding of the semantic scope required by the prompt, often retrieving information that, while contextually related, diverged significantly from the specific data fields defined by the taxonomy (e.g., returning descriptive actions instead of categorical labels like profession or relationship type).

One of the primary limitations encountered was the ambiguity in natural language and its impact on the LLMs' reasoning process. This was especially evident when models were asked to infer information indirectly stated or entirely absent from the text. Instead of indicating the lack of evidence, models frequently hallucinated responses, fabricating plausible but unfounded details. This behavior critically undermines the reliability of extracted data, particularly in legally sensitive contexts.

Another noteworthy limitation was the tendency of models to prioritize certain lexical or structural cues over deeper contextual understanding. This resulted in erroneous classification of attributes such as gender, age, and relationship roles, particularly in complex or non-standardized sentence structures. Furthermore, despite clear instructions embedded in the prompt (e.g., limiting response length or choosing from a set of predefined options), the outputs regularly violated these constraints by

including a rationale that justifies the provided answer, revealing the models' limited capacity for controlled generation. Nevertheless, such an explanation is not only not requested, but is also frequently illogical or based on spurious correlations, thereby accentuating the interpretability issue.

The comparison of the selected prompts demonstrated that the adoption of direct instruction prompts, which explicitly instructed the model to select from provided options or adhere to strict syntactic patterns[5], resulted in a substantial enhancement in performance stability. Nevertheless, the more general limitations in comprehension and factual accuracy persist, particularly in circumstances where information is partial or ambiguous.

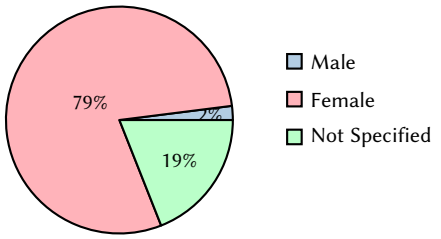
4.2. Extracted Statistics

The statistical analysis was carried out on a set of 607 anonymized judicial rulings. This final number resulted from a filtering process that excluded rulings exceeding the token limits of the models used, as well as those containing errors introduced during the OCR extraction of the original documents. After applying these cleaning steps, 607 out of the original 865 rulings were deemed suitable for analysis.

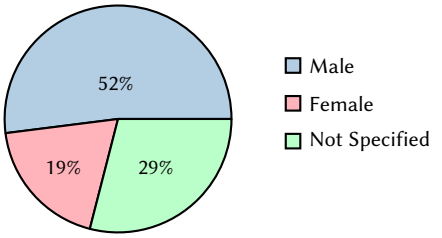
As discussed in Section 3.1, we focus on the results obtained from the best-performing model, Phi-3-Mini (4B), which demonstrated strong performance while maintaining low computational requirements. All generations are produced using greedy decoding to allow reproducibility, with the maximum number of tokens set to 512. The extraction process was guided by the adoption of the prompts selected in the prompt evaluation phase, with the objective of capturing relevant characteristics and extracting statistics and trends that would encompass the entire taxonomy area.

Demographic Trends A significant skew emerged in the gender distribution of both victims and culprits. As shown in Figure 1a, the inferred victims were predominantly female, comprising approximately 79% of the identified cases. In contrast, as shown in figure 1b, the majority of culprits were male, accounting for 52% of the dataset. These figures align with established criminological patterns observed in domestic and gender-based violence cases. A notable proportion of records (19% for victims and 29% for perpetrators) lacked sufficient information to determine gender, reflecting the limitations imposed by anonymization and the challenges in automatic extraction.

[5]As an example when asking for the victim gender: *Qual è il genere della vittima? Rispondi con "maschio", "femmina" o "non specificato"* which translates to *What is the victim gender? Reply with "male", "female" or "not specified".*



(a) Pie Chart representing the victims' gender distribution.



(b) Pie Chart representing the culprits' gender distribution.

Figure 1: Gender distribution of victims and culprits.

A similar phenomenon was observed in the data pertaining to nationality. The majority of individuals identified as both victims and culprits were of Italian origin (89% and 90% respectively). A mere proportion of the subjects belonged to minority groups, with Nigerian, Chinese, and Albanian nationals being the most frequently mentioned among non-Italian individuals. In some cases (1,3% and 2,1% for culprits and victims), the nationality of the subjects could not be established due to the absence of explicit references within the anonymised texts.

Nature of Relationships A thorough analysis of interpersonal relationships indicated that the majority of crimes occurred within familiar or intimate settings. As represented in Figure 2, conjugal relationships were the most frequently identified type of relationship (over 30% of cases), followed closely by cohabiting arrangements (over 21% of cases). These findings underscore the imperative for meticulous examination of domestic environments as pivotal contexts for violent offences. A small yet noteworthy proportion of cases (around 2% of cases) exhibited ambiguous or non-identifiable relationships, thereby further emphasising the complexity involved in disambiguating personal information within anonymised legal documents, which frequently report such information in an indirect form.

Crime Scene and Modus Operandi The most frequent locations linked to criminal acts were private res-

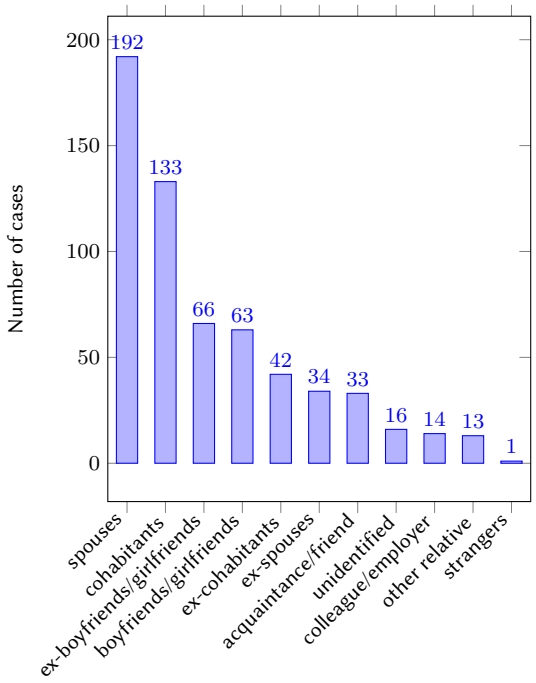


Figure 2: Relationship between the victim and the culprit in analyzed cases.

idences (approximately 47%), with a breakdown of 10% occurring in the victim's residence, 15% in the perpetrator's residence, and 22% in other residences not belonging to either party. Open public spaces accounted for over 18% of cases. In 13% of cases, the location of the crime could not be determined based on the available information. The remaining proportion comprises the other locations outlined in the taxonomy.

With regard to the weapons involved in the crime, as shown in Figure 3, approximately half of the records indicated that no identifiable instrument was present. This is indicative of both non-violent offences and limitations in the reporting or modelling process. Among the detected weapons, the most prevalent are firearms (21% of the cases) and blunt objects (15% of the cases). These distributions are consistent with the high frequency of lethal or severely injurious outcomes reported in the corpus.

Typologies of Crime and Motivation The most prevalent offence detected within the corpus is homicide (around 36% of the cases), constituting over one-third of all analysed court rulings. Other prevalent categories included personal injury, physical assault, and threats (12%, 9% and 7% respectively), which often co-occur with domestic or interpersonal conflict. Finally, in terms of motive, quarrels/futile motives, insanity and grudges (38%,

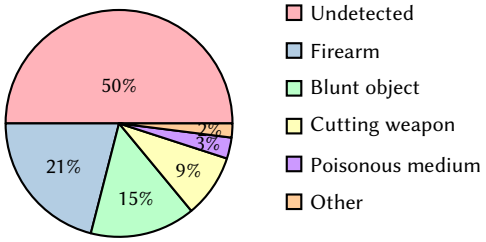


Figure 3: Pie chart representing the relative distribution of weapons used to commit the crime.

24.2%, and 23.9% respectively) emerge as most frequent.

5. Conclusions

The frequent occurrence of missing or indeterminable values across multiple dimensions, such as gender, nationality and location, highlights a structural limitation when working with anonymised legal texts. Furthermore, reliance on automatic extraction tools introduces additional uncertainty, particularly in complex or syntactically ambiguous contexts.

The prompt selection phase underscored a fundamental tension between the expressive power of LLMs and their reliability in high-precision tasks. While the models demonstrated potential in handling straightforward cases, their performance deteriorated significantly in edge cases or when faced with incomplete data.

Nevertheless, statistics extracted using carefully selected prompts provide a compelling insight into the sociological and criminological patterns embedded in the Italian judicial landscape. These statistics demonstrate the potential of language models in supporting data-driven legal analysis. However, they also reveal the need for enhanced model guidance, human oversight and methodological rigor to ensure the validity of the insights produced.

A promising direction for future work involves conducting a systematic evaluation using human-annotated data to more rigorously assess the model's accuracy and reliability in extracting structured information from legal texts.

Acknowledgments

The work of Daniel Scalena has been partially funded by MUR under the grant ReGAInS, Dipartimenti di Eccellenza 2023-2027 of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca.

References

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, ACM Transactions on Intelligent Systems and Technology (2023).

[2] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3816–3830. URL: https://aclanthology.org/2021.acl-long.295/. doi:10.18653/v1/2021.acl-long.295.

[3] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3?, in: E. Agirre, M. Apidianaki, I. Vulić (Eds.), Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Association for Computational Linguistics, Dublin, Ireland and Online, 2022, pp. 100–114. URL: https://aclanthology.org/2022.deelio-1.10/. doi:10.18653/v1/2022.deelio-1.10.

[4] L. Wang, N. Yang, F. Wei, Learning to retrieve in-context examples for large language models, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1752–1767. URL: https://aclanthology.org/2024.eacl-long.105/.

[5] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 50–70. URL: http://dx.doi.org/10.1109/TKDE.2020.2981314. doi:10.1109/tkde.2020.2981314.

[6] I. Chalkidis, I. Androutsopoulos, A. Michos, Extracting contract elements, in: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, ICAIL '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 19–28. URL: https://doi.org/10.1145/3086512.3086515. doi:10.1145/3086512.3086515.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261/. doi:10.18653/v1/2020.findings-emnlp.261.

[8] D. Mamakas, P. Tsotsi, I. Androutsopoulos, I. Chalkidis, Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer, in: N. Aletras, I. Chalkidis, L. Barrett, C. Goanță, D. Preoțiuc-Pietro (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 130–142. URL: https://aclanthology.org/2022.nllp-1.11/. doi:10.18653/v1/2022.nllp-1.11.

[9] D. Mali, R. Mali, C. Barale, Information extraction for planning court cases, in: N. Aletras, I. Chalkidis, L. Barrett, C. Goanță, D. Preoțiuc-Pietro, G. Spanakis (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2024, Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 97–114. URL: https://aclanthology.org/2024.nllp-1.8/. doi:10.18653/v1/2024.nllp-1.8.

[10] C. Barale, M. Rovatsos, N. Bhuta, Automated refugee case analysis: An NLP pipeline for supporting legal practitioners, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2992–3005. URL: https://aclanthology.org/2023.findings-acl.187/. doi:10.18653/v1/2023.findings-acl.187.

[11] M. Cemri, T. Çukur, A. Koç, Unsupervised simplification of legal texts, 2022. URL: https://arxiv.org/abs/2209.00557. arXiv:2209.00557.

[12] J. Zhao, Y. Wang, N. Rusnachenko, H. Liang, Legal_try at SemEval-2023 task 6: Voting heterogeneous models for entities identification in legal documents, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1282–1286. URL: https://aclanthology.org/2023.semeval-1.178/. doi:10.18653/v1/2023.semeval-1.178.

[13] J. Niklaus, V. Matoshi, M. Stürmer, I. Chalkidis, D. Ho, MultiLegalPile: A 689GB multilingual legal corpus, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15077–15094. URL: https://aclanthology.org/2024.acl-long.805/. doi:10.18653/v1/2024.acl-long.805.

[14] M. Masala, R. C. A. Iacob, A. S. Uban, M. Cidota, H. Velicu, T. Rebedea, M. Popescu, jurBERT: A

Romanian BERT model for legal judgement prediction, in: N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, D. Preotiuc-Pietro (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 86–94. URL: https://aclanthology.org/2021.nllp-1.8/. doi:10.18653/v1/2021.nllp-1.8.

[15] M. Rovera, A. Palmero Aprosio, F. Greco, M. Lucchese, S. Tonelli, A. Antetomaso, Italian legislative text classification for gazzetta ufficiale, in: D. Preoțiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. Spanakis, N. Aletras (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2023, Association for Computational Linguistics, Singapore, 2023, pp. 44–50. URL: https://aclanthology.org/2023.nllp-1.6/. doi:10.18653/v1/2023.nllp-1.6.

[16] D. Licari, G. Comandè, ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law, in: D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. D. Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, N. Troquard (Eds.), Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, volume 3256 of *CEUR Workshop Proceedings*, CEUR, Bozen-Bolzano, Italy, 2022. URL: https://ceur-ws.org/Vol-3256/#km4law3, iSSN: 1613-0073.

[17] L. Team, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[18] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024.

A. Appendix - Taxonomy

Figure 4 reports a schematic representation of the taxonomy developed to model the relationships relevant to the definition of offences and associated entities. The taxonomy integrates and reorganises classifications provided by the Istituto Nazionale di Statistica (ISTAT) to ensure a comprehensive and valid characterization of the analysed legal court rulings.

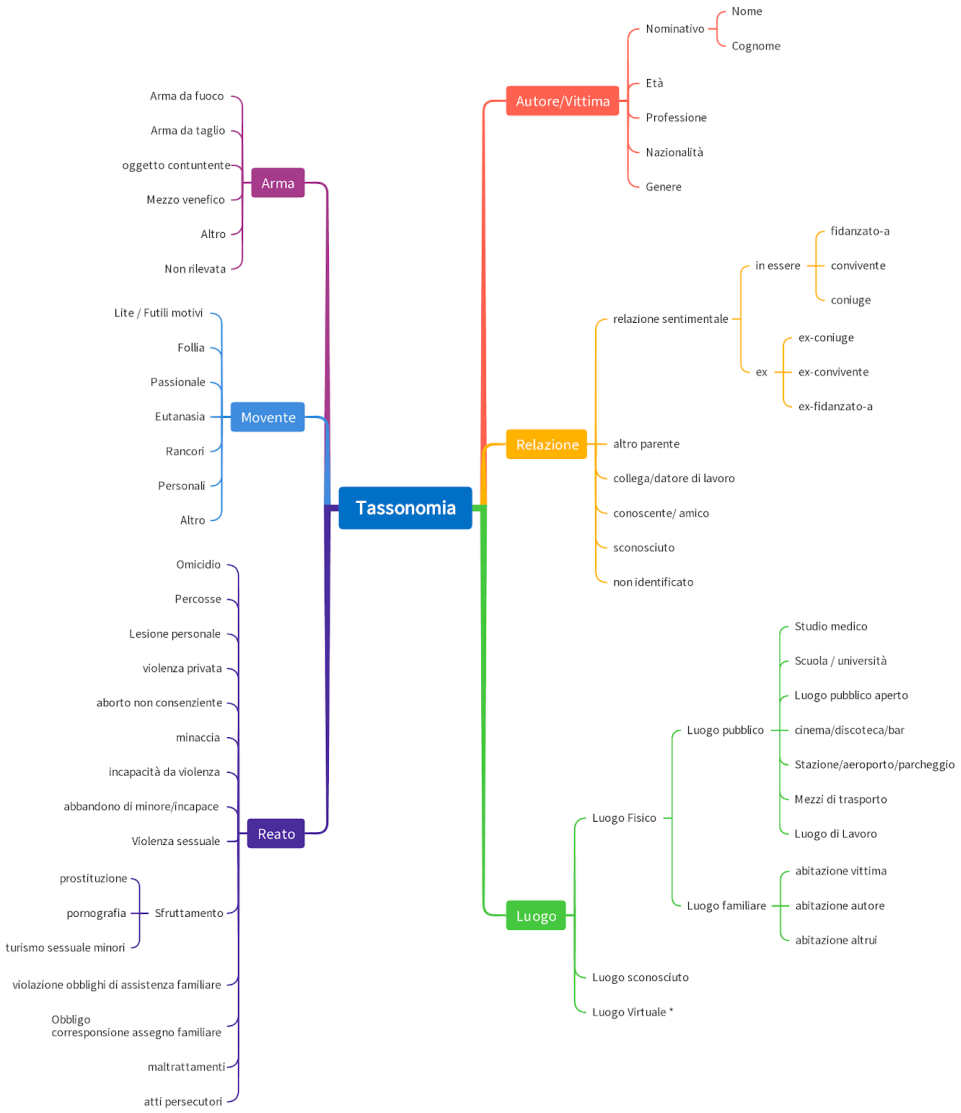


Figure 4: Taxonomy modeling key relationships for offence definition and entity identification, based on ISTAT classifications.