A Methodology for Extracting Key Information from Crime News

Eleonora Calò[1,*], Loredana Caruccio[1], Genoveffa Tortora[1] and Livio Vona[1]

[1] *Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy*

Abstract

Official crime statistics are often released with significant delays and in an aggregated form that limits their usefulness for real-time monitoring and detailed analysis. In different countries, as happens in Italy, accessing to detailed crime records is further restricted by the confidential nature of police reports. Conversely, online news articles offer an alternative source of information that is rich and timely. However, their unstructured textual content poses challenges for automated data extraction processing. In this paper, we present a Natural Language Processing (NLP) pipeline designed to extract structured information from Italian crime news articles, which has been evaluated on a semi-automatically annotated dataset. Despite the limited amount of annotated data, the proposed approach showed promising results: the topic modeling component, included into the pipeline, facilitated targeted crime analysis, and the Named Entity Recognition (NER) module achieved an overall F1-score of 74%. These results demonstrate the effectiveness of the methodology in low-resource settings and highlight its potential for supporting automated crime analysis based on unstructured news data.

Keywords

Topic extraction, Crime News, Natural Language Processing

1. Introduction

Crime is a complex phenomenon, and the ways in which it is addressed significantly vary depending on the legal and institutional framework of each country. In particular, the processes involved in collecting, managing and disseminating crime-related information can widely differ from one country to another. While many governments publish official statistics on the location, date and type of crimes, this data is often released with considerable delays, which limits timely monitoring. Furthermore, richer, more detailed sources, such as police records, usually remain confidential and are not publicly accessible.

This is the case in Italy, where the only reliable data source in this domain is the National Institute of Statistics (ISTAT)[1]. However, the information it provides is aggregated and typically becomes publicly available only at least one year after the incidents have occurred. This highlights the need to explore alternative sources of information that can complement or, in some cases, anticipate the evidence provided by institutional channels. One example of such a source is online crime news articles, which are often published in a timely manner when a crime occurs. This can help to reduce the information gap. Nevertheless, although Natural Language Processing (NLP) techniques make it possible to extract specific, detailed information about crimes, dealing with textual corpora is always difficult due to ambiguous sentences and different writing styles. It is also challenging to guarantee a proper characterization of domain knowledge.

To address this need, we developed a pipeline consisting of three main steps: web scraping on Italian crime news articles, crime categorization using BERTopic [1], and finally information extraction using Named Entity Recognition (NER). Through the developed pipeline, it has been possible to identify crimes in which weapons were used. On this specific category of crime, the application of BERT-based

1st Workshop on supporting CRIme reSolution Through Artificial INtelligence (CRISTAIN), held in conjunction with CHITALY 2025, Salerno, Italy, October 6–10

[*] Corresponding author.

[†] All authors contributed equally.

ecalo@unisa.it (E. Calò); lcaruccio@unisa.it (L. Caruccio); tortora@unisa.it (G. Tortora); l.vona1@studenti.unisa.it (L. Vona)

[1] https://www.istat.it/

models and NLP techniques enabled the automatic extraction of detailed information directly from the text of the articles, such as the identity of the perpetrator, the victim involved, the type of weapon used, and other contextual elements. This approach allows for a more complete and structured picture of the different crimes occurred on the Italian territory, making possible having more in-depth and targeted analysis. In addition, the information extracted can be exploited to feed monitoring systems, build knowledge bases, support the investigation or develop decision-support tools in institutional settings.

From this pipeline and the obtained results, this paper aims to investigate some key questions allowing the evaluation of the effectiveness of the proposed approach. Specifically, in this paper we try to provide answers to the following Research Questions (RQs):

RQ1: What is the most compliant evaluation scheme to use in the context of crime?
RQ2: How effective is the model with respect to the different categories of entities?
RQ3: What types of errors emerge in the identification of individual entities?

The remainder of the paper is organized as follows: Section 2 discusses different work related to the extraction of key information from crime cases; Section 3 describes the pipeline and its various stages; Section 4 discusses the performed experimental evaluation phase; and finally, Section 5 outlines conclusions about the work done and possible future developments.

2. Related work

This section addresses the problem of extracting key information from crime-related cases and explores the most common approaches used in this domain. Based on a review of the literature, three main categories of approaches are identified:

- **Rule-based approaches** that rely on predefined syntactic and semantic rules to identify and classify information. They are accurate but lack adaptability to new or unexpected contexts.
- **Machine Learning (ML) based approaches** that use statistical models trained on annotated data to recognize patterns and structures. They are more flexible than rule-based systems but require large amounts of labeled data.
- **Large Language Model (LLM) based approaches** that leverage pre-trained language models (e.g., GPT or BERT) to understand and generate complex text. They are highly effective with unstructured data and complex scenarios.

Rule-based Approaches. These methods are effective for extracting basic information such as location, date, and names from crime news articles. Rahma et al. [2] combined dependency parsing and POS-tagging to extract metadata (e.g., crime type, victim, criminal) from Indonesian crime articles, achieving an F1-score of 60%. Dharviyanti et al. [3] used rule-based extraction for four entities: date, crime type, location, and persons, achieving an F1-score of around 90%.

Although precise, these approaches are limited in generalizing across linguistic patterns and require expert knowledge for the definition of rule.

Machine Learning Approaches. ML-based techniques often deal with the task as NER. Arulanandam et al. [4] used Conditional Random Fields to identify crime-related sentences and applied NER to extract theft locations. Dasgupta et al. [5] used NER and Support Vector Machine (SVM) classifiers to extract metadata like names, dates, crime types, and actions. Sedik et al. [6] used SVM to detect crime-related sentences in Indonesian texts and applied NER to extract date and location.

These approaches offer flexibility but struggle with context-dependent information and heavily rely on large annotated datasets.

LLM-Based Approaches. These methods offer greater flexibility and generalization. Their subtypes include Sequence Labeling, Extractive Question Answering (QA), and Instruction-Tuned Models.
In *Sequence Labeling*, each token in the input sequence is assigned to a label, enabling the model to identify and extract structured information, such as named entities or metadata fields. Pongpaichet et al. [7] used BIO-tagging and fine-tuned transformer models like WangchanBERTa and XLM-R on 4,000 Thai crime articles. They achieved an F1-score of 67% for closed-domain metadata extraction.
In the *Extractive QA* method, the model selects text segments that answer predefined questions based on the input document. Rollo et al. [8] employed a BERT-based QA system using the 5W+1H (What, When, Where, Who, Why, How) framework to extract events from Italian crime articles and represent them in a knowledge graph. Cao et al. [9] created a manually annotated 5W+1H dataset from four news corpora and evaluated different LLMs using ROUGE and BLEU metrics [10]. Fine-tuned models outperformed ChatGPT and approached GPT-4's few-shot performance.
An *Instruction-Tuned Model* is a LLM that has been further trained on a dataset of instructions and corresponding outputs in order to improve its ability to follow instructions and perform specific tasks. Park et al. [11] used GPT-3.5 for prompt engineering to generate a labeled Korean crime dataset, then several fine-tuned models (e.g., KLUE-BERT, XLM-RoBERTa) use both token and sequence classification. They achieved an F1-score of 87%.

Italian Language Approaches. Although most of the studies are not focused on Italian, as described above, similar approaches have been used in other domains, such as the medical one. In fact, Crema et al. [12] developed the NLP Extraction and Management Tool (NEMT), a semi-automatic tool based on a QA bot fine-tuned on clinical QA datasets, achieving an F1-score of 84.7%. Buonocore et al. [13] proposed a BERT-based NER system to extract events from cardiology reports, using an IOB2-tagged dataset. They also included an extractive QA component for clinical registry tasks.

Nevertheless, there is the need of exploiting (semi-)automatic pipelines capable of properly analyze texts, extract information related to the specific domain of crime, and possibly relate the information extraction step with other proper variables, as in the case of the crime category. The proposed methodology represent one of the first strategies following this direction.

3. Methodology

The proposed approach is based on an NLP pipeline, developed to extract structured information from Italian crime-related news articles (see Figure 1). The system is organized into three main phases:

1. **Data collection and processing (Web scraping)** that uses libraries like Requests[2] and BeautifoulSoup[3] to fetch and parse articles in order to extract useful information like the main text;
2. **Topic Identification (Categorization)** that uses the BERTopic model, trained on the input articles in order to generate different crimes categories; and
3. **Information Extraction** that, as main phase, aims to annotate the dataset for training of different BERT models and to extract domain related entities (i.e., through a NER process) through the spaCy[4] library.

For demonstration purposes, the presented pipeline is customized on crime articles involving weapons.

3.1. Data collection and processing

Due to the lack of official and timely crime data, we considered online crime news, collecting them to compose a dataset that consists of roughly $4,500$ of italian articles published between January 2023

[2] https://requests.readthedocs.io/en/latest/
[3] https://beautiful-soup-4.readthedocs.io/en/latest/
[4] https://spacy.io/

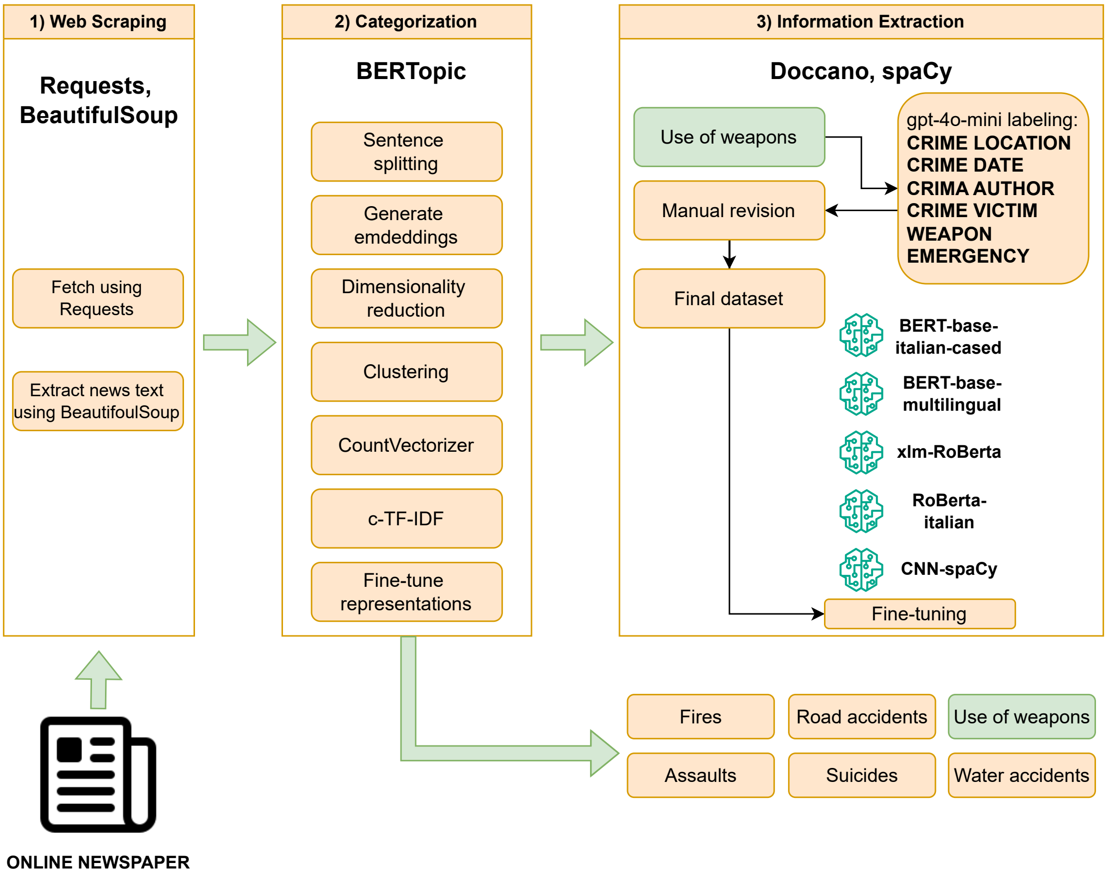


Figure 1: The proposed pipeline for the web scraping (1), categorization (2), and extraction (3) of key information from Italian crime news.

and February 2025. The source of online newspaper was Virgilio Notizie website[5] (see Figure 1 - Web scraping). The collected newspaper were processed to extract meaningful information, such as *title, publication date, description, url,* and *main text*. All these information was then saved in a JSON format for further analysis.

3.2. Topic Identification through BERTopic

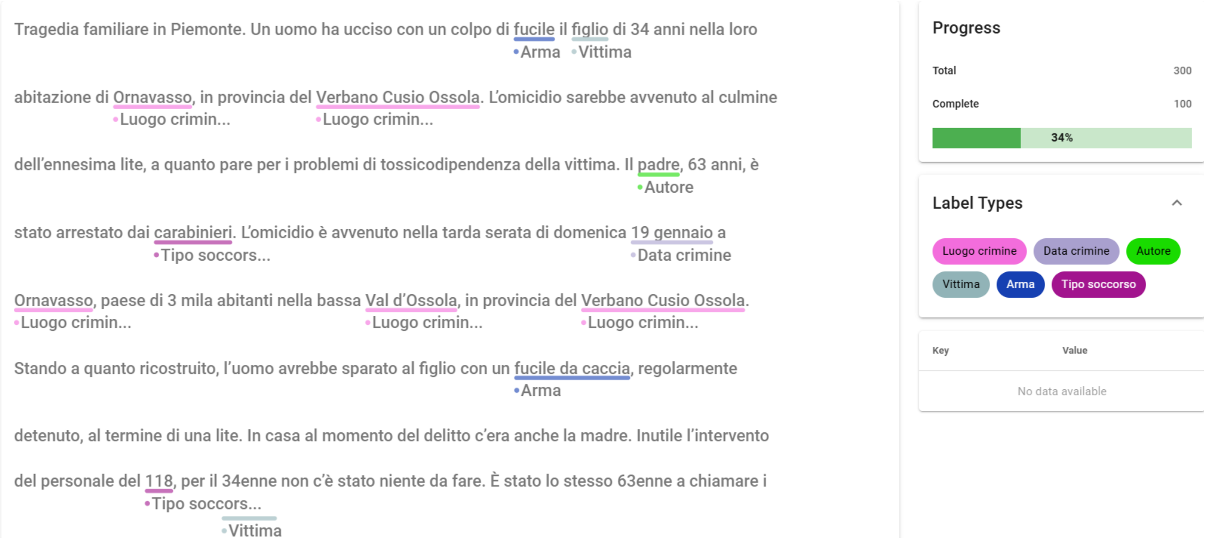
To group articles according to their content and categorize them in different crime categories, we employed BERTopic [1], a topic modeling technique that leverages sentence embeddings and unsupervised clustering. Embeddings were generated using a multilingual Sentence-BERT[6] model, then reduced in dimensionality using UMAP [14]. These embeddings were subsequently clustered using HDBSCAN [15], which does not require a predefined number of clusters and can identify possible noise.

Once the clusters were defined, the most representative terms for each topic were extracted using the c-TF-IDF algorithm [1]. To enhance the interpretability of these topics, we also applied KeyBERT[7], Maximal Marginal Relevance[8] and GPT-4o-mini[9] in zero-shot mode, allowing to assign meaningful labels. As shown in Figure 1, the output of the Categorization phase, consists of different possible crime categories, such as crimes involving *use of weapons, fires, assaults, suicides,* and *road accidents*.

[5] https://www.virgilio.it/notizie/cronaca/cronaca-nera/
[6] https://sbert.net/
[7] https://maartengr.github.io/BERTopic/getting_started/representation/representation.html#keybertinspired
[8] https://maartengr.github.io/BERTopic/getting_started/representation/representation.html#maximalmarginalrelevance
[9] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

3.3. Information Extraction with Named Entity Recognition

As shown in Figure 1, in the final phase, i.e., Information Extraction, the goal is to extract six specific types of entities from the crimes involving the use of *weapons*. The extracted entities are: *author, victim, used weapon, location and date of the crime, and any involved emergency services.* Since no annotated dataset for this domain was readily available, we adopted a semi-automatic annotation strategy. Initial annotations were produced with GPT-4o-mini and then manually reviewed and corrected by using the Doccano annotation tool[10] (see Figure 2).



Tragedia familiare in Piemonte. Un uomo ha ucciso con un colpo di fucile il figlio di 34 anni nella loro
abitazione di Ornavasso, in provincia del Verbano Cusio Ossola. L’omicidio sarebbe avvenuto al culmine
dell’ennesima lite, a quanto pare per i problemi di tossicodipendenza della vittima. Il padre, 63 anni, è
stato arrestato dai carabinieri. L’omicidio è avvenuto nella tarda serata di domenica 19 gennaio a
Ornavasso, paese di 3 mila abitanti nella bassa Val d’Ossola, in provincia del Verbano Cusio Ossola.
Stando a quanto ricostruito, l’uomo avrebbe sparato al figlio con un fucile da caccia, regolarmente
detenuto, al termine di una lite. In casa al momento del delitto c’era anche la madre. Inutile l’intervento
del personale del 118, per il 34enne non c’è stato niente da fare. È stato lo stesso 63enne a chiamare i

Figure 2: A Doccano snapshot of the labeling process for the crime entity extraction task.

4. Experimental Evaluation

This Section presents the experiments performed on the different NER models evaluated in the proposed approach. The evaluation is organized according to three research questions (RQ1–RQ3). RQ1 investigates which evaluation schema best captures performance in this domain, recognizing that traditional metrics at token level may not be the best approach. RQ2 explores the model’s performance across different entity types to identify strengths and weaknesses in recognizing specific categories. Finally, RQ3 examines the nature and sources of detected errors, aiming to uncover persistent issues, such as class imbalance and context ambiguity that impact the model accuracy.

Technical Settings. We collected and semi-automatically annotated a dataset of around 100 samples, which has been used to fine-tune different BERT-based models on a NER task. The models are: BERT-base-italian-cased[11], BERT-base-multilingual[12], xlm-RoBerta[13], RoBerta-italian[14] and CNN-spaCy model. All models were fine-tuned for up to 80 epochs, with a learning rate of $5 \times (10^{-5})$, using Adam as optimizer, while the smart batching technique was used with a dynamic batch size. In addition, a strided-spans technique was used to handle texts that exceed the BERT context limit.

Preliminary results. To evaluate and select the best model, we used the F1-score metric on the results obtained from the validation set. As shown in Figure 3, the BERT-base-italian-cased model demonstrated

[10] https://github.com/doccano/doccano
[11] https://huggingface.co/dbmdz/bert-base-italian-cased
[12] https://huggingface.co/google-bert/bert-base-multilingual-uncased
[13] https://huggingface.co/FacebookAI/xlm-roberta-base
[14] https://huggingface.co/osiria/roberta-base-italian

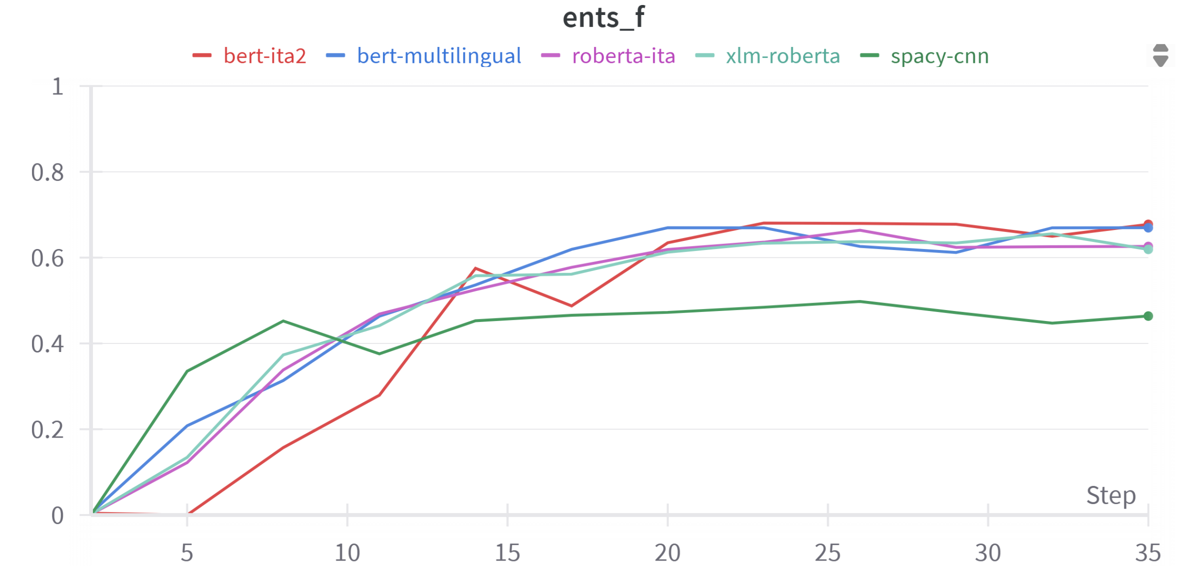


Figure 3: F1-score values on validation set.

the best performance compared to the other models, with an F1-score of 68%. Subsequently, the selected model was evaluated on an unseen test set to assess its generalization capability, achieving a strict F1-score of 65% (see Table 1).

RQ1. What is the most suitable evaluation schema for the crime domain? To identify the most appropriate scoring strategy, for evaluation schemas from SemEval 2013 were compared (see Table 1):

- *strict:* exact string match and entity type;
- *exact:* exact string match regardless of the type;
- *partial:* partial string match regardless of the type;
- *ent_type:* some overlap between the system tagged entity and the gold annotation.

The results indicated that the *ent_type* schema is the most appropriate in the context of crime-related entity extraction, as it accounts for entity type correctness even when the span is not perfectly matched (Table 2 shows several examples of unrecognized matching through the *strict* evaluation schema). In fact, using the *ent_type* schema, the best model (i.e., BERT-base-italian-cased) achieved an overall F1-score of 74%, compared to 65% with the *strict* schema.

RQ2. How effective is the model with respect to the different categories of entities? The performance analysis was conducted per entity type (see Figure 4). The model showed high performance on specific categories, especially for *weapon* (F1-score: 93%) and *crime date* (F1-score: 87%). Intermediate performance was observed for *emergency response* (F1-score: 79%), *victim* (F1-score: 70%), and *crime location* (F1-score: 69%). The lowest result was obtained for the *author* entity, which reached only 43%, suggesting difficulties in detecting semantically ambiguous or context-dependent references to the author of the crime.

	correct	incorrect	partial	missed	spurious	Precision	Recall	F1-score
ent_type	116	4	0	29	43	0.71	0.78	0.74
partial	103	0	17	29	43	0.68	0.75	0.71
strict	101	19	0	29	43	0.62	0.68	0.65
exact	103	17	0	29	43	0.63	0.69	0.66

Table 1
Results for different evaluation schemas as defined in SemEval 2013.

true entity	true span	predicted entity	predicted span
Author	uomo di 39 ann	Author	uomo di 39 anni
Crime location	parcheggio del supermercato	Crime location	parcheggio del supermercato Lidl
Emergency services	sanitari del 118	Emergency services	118
Crime location	Sannicandro di Bari	Crime location	Sannicandro
Author	78enne	Author	un 78enne
Author	45enne albanese	Author	45enne
Crime date	lunedì 8 aprile	Crime date	8 aprile
Crime date	Venerdì 17 maggio	Crime date	17 maggio

Table 2
Incorrect type error analysis for the strict evaluation schema.

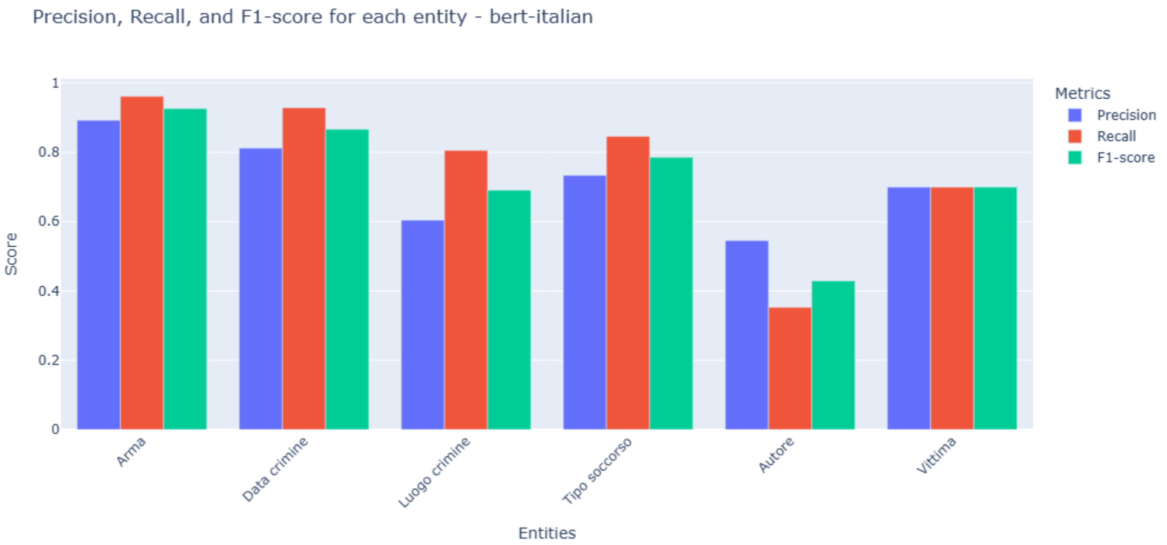


Figure 4: Metrics per entity-type following the **ent_type** schema.

RQ3. What types of errors emerge in the entity recognition process? As shown in Table 3, an analysis process concerning the model-prediction errors revealed that missed entities (i.e., false negatives) for the *author* class and spurious entities (i.e., false positives) for the *crime location* class. This behavior is due to the lack of representative examples in the training data, particularly for the *author* class, which negatively impacted the model generalization. On the other hand, many false positives for the *crime location* entity were caused by the presence of secondary locations in the articles. These observations indicate that class imbalance and entity context disambiguation remain key challenges to address in the future.

Discussion. The NER module demonstrated good generalization capabilities on some specific entities, such as *weapon* (F1-score 93%) and *crime data* (F1-score 87%), while it encountered greater difficulties on more ambiguous entities such as *author* (F1-score 43%), where semantic ambiguity, small errors in the annotation phase and the scarcity of representative examples had a negative impact on performance.

	correct	incorrect	partial	missed	spurious	Precision	Recall	F1-score
Weapon	25	0	0	1	3	0.89	0.96	**0.93**
Author	6	1	0	**10**	4	0.55	0.35	**0.43**
Crime date	13	0	0	1	3	0.81	0.93	**0.87**
Crime location	29	1	0	6	**18**	0.60	0.81	**0.69**
Emergency services	22	0	0	4	8	0.73	0.85	**0.79**
Victim	21	2	0	7	7	0.70	0.70	**0.70**

Table 3
Result distributions and performance values per entity-type following the **ent_type** schema.

These results confirm both the validity of the adopted approach and the crucial importance of the quality of the annotated dataset.

5. Conclusion and Future Directions

In this paper, we propose an automated methodology for extracting key concepts from Italian-language crime news articles. The aim is to fill the information gaps left by institutional sources, which are often aggregated or accessed with a delay. To this end, we developed a pipeline consisting of three main modules: web scraping for collecting online news, categorizing news using topic modelling (BERTopic) and extracting structured entities through NER, based on fine-tuned BERT models. Despite the limited availability of annotated data (about 100 samples), the overall results were promising, with high performance on well-defined entities, such as *weapon* and *data crime*, and the presence of a few critical entities to recognize, such as the *author*. This confirms the effectiveness of the approach and highlights the importance of the quality of annotated datasets.

In the future, we would like to integrate the descibed NLP pipeline into a tool. It would allow the automatic identification of entities within texts and/or articles. Then, through a user parameterization schema, it would be possible to select the entities of interest and automatically generate structured datasets in order to be used to conduct further analyses in the domain of crime.

Acknowledgments

This study was funded by Ministero dell'Università e della Ricerca (MUR) of Italy in the context of project denoted as BLOODSTAIN in the program PRIN 2022 (grant number D53D23008660006).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. Grootendorst, Bertopic: Neural topic modeling with a class-based TF-IDF procedure, CoRR (2022).

[2] F. Rahma, A. Romadhony, Rule-based crime information extraction on indonesian digital news, in: International Conference on Data Science and Its Applications (ICoDSA), 2021, pp. 10–15.

[3] N. A. D. Dharviyanti, N. Wilantika, Rule-based ner for crime information extraction through online news site, in: International Conference on Information Technology Research and Innovation (ICITRI), 2024, pp. 99–104.

[4] R. Arulanandam, B. T. R. Savarimuthu, M. A. Purvis, Extracting crime information from online newspaper articles, in: Proceedings of the second australasian web conference-volume 155, volume 155, 2014, pp. 31–38.

[5] T. Dasgupta, A. Naskar, R. Saha, L. Dey, Crimeprofiler: Crime information extraction and visualization from news media, in: Proceedings of the international conference on web intelligence, 2017, pp. 541–549.

[6] R. R. Sedik, A. Romadhony, Information extraction from indonesian crime news with named entity recognition, in: 15th International Conference on Knowledge and Smart Technology (KST), IEEE, 2023, pp. 1–5.

[7] S. Pongpaichet, B. Sukosit, C. Duangtanawat, J. Jamjongdamrongkit, C. Mahacharoensuk, K. Matangkarat, P. Singhajan, T. Noraset, S. Tuarob, Camelon: A system for crime metadata extraction and spatiotemporal visualization from online news articles, IEEE Access 12 (2024) 22778–22802.

[8] F. Rollo, L. Po, G. Bonisoli, et al., Online news event extraction for crime analysis., in: Italian Symposium on Advanced Database Systems (SEBD), 2022, pp. 223–230.

[9] Y. Cao, Y. Lan, F. Zhai, P. Li, 5w1h extraction with large language models, in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2024, pp. 1–8.

[10] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 7881–7892.

[11] Y. Park, R. S. Park, H. Kim, Key information extraction for crime investigation by hybrid classification model, Electronics 13 (2024) 1525.

[12] C. Crema, F. Verde, P. Tiraboschi, C. Marra, A. Arighi, S. Fostinelli, G. M. Giuffré, V. P. Dal Maschio, F. L'abbate, F. Solca, et al., Medical information extraction with nlp-powered qabots: A real-world scenario, IEEE Journal of Biomedical and Health Informatics 28 (2024) 6906–6917.

[13] T. M. Buonocore, E. Parimbelli, V. Tibollo, C. Napolitano, S. Priori, R. Bellazzi, A rule-free approach for cardiological registry filling from italian clinical notes with question answering transformers, in: International Conference on Artificial Intelligence in Medicine, Springer, 2023, pp. 153–162.

[14] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv (2018).

[15] L. McInnes, J. Healy, S. Astels, et al., hdbscan: Hierarchical density based clustering., J. Open Source Softw. 2 (2017) 205.