# DICE: A Dataset of Italian Crime Event News

Giovanni Bonisoli, Maria Pia Di Buono, Laura Po, Federica Rollo

University of Modena and Reggio Emilia, Modena, Italy
University of Napoli "L'Orientale" Napoli, Italy
Contact: **giovanni.bonisoli@unimore.it**

## Background and contributions

The **information extraction from news articles** is the task of automatically acquiring useful information from news articles, to allow for more efficient processing, retrieval, and analysis of massive news data that is nowadays available in the digital form.

Our main contributions are:

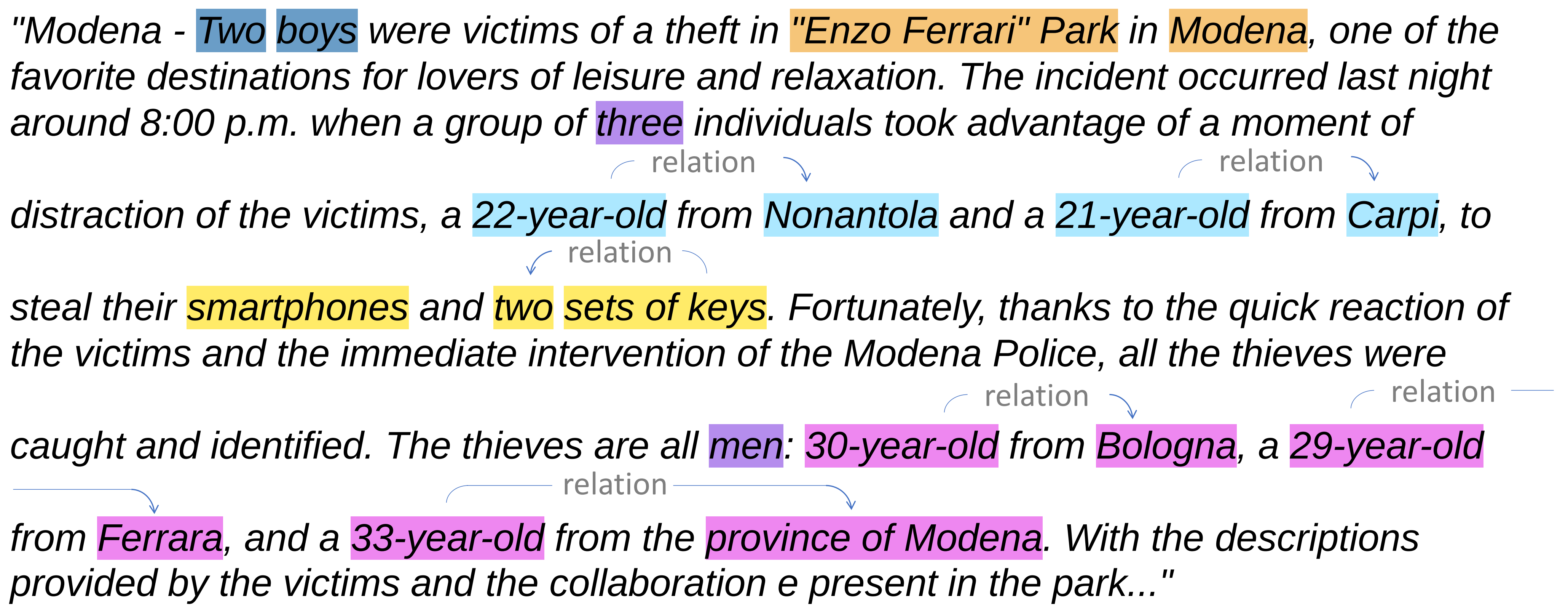
1. a **Dataset for Italian Crime Event News** (DICE), containing **10,395 crime news** enriched with an automatic annotation.
2. a **new annotation schema** to manually extract the main information about crime events from news text.
3. a preliminary manual annotation using the proposed schema of **1000 documents**.

## Dataset

| | |
|---|---:|
| News selected | 10,395 |
| Geolocalized news | 8,295 |
| NER objects | 75,256 |
| Dbpedia link | 42,545 |
| Time expression | 20,832 |
| Manually annotated news | 1000 |
| Single event theft news | 406 |

## Annotation schema

| Label | Description | Relations |
|---|---|---|
| LOC | The location of a crime | No |
| VIC | The experiencer of a crime | Yes |
| VICG | A group of experiencers | No |
| AUT | The causer of a crime | Yes |
| AUTG | A group of causers | No |
| OBJ | Robbed object(s) | Yes |
| PAR | An injured party in a crime | No |


*"Modena - Two boys were victims of a theft in "Enzo Ferrari" Park in Modena, one of the favorite destinations for lovers of leisure and relaxation. The incident occurred last night around 8:00 p.m. when a group of three individuals took advantage of a moment of distraction of the victims, a 22-year-old from Nonantola and a 21-year-old from Carpi, to steal their smartphones and two sets of keys. Fortunately, thanks to the quick reaction of the victims and the immediate intervention of the Modena Police, all the thieves were caught and identified. The thieves are all men: 30-year-old from Bologna, a 29-year-old from Ferrara, and a 33-year-old from the province of Modena. With the descriptions provided by the victims and the collaboration e present in the park..."*

## Annotation process

- **2 rounds** of annotation to calibrate guidelines.
- 3 expert annotators and 1 non-expert annotator.
- Measurement of the **Inter-Annotator Agreement** (IAA) through the **Krippendorff's** $\alpha$.

| | IAA 1° round | IAA 2° round |
|---|---|---|
| **AUT** | 0.605 | **0.85** |
| **AUTG** | n.a. | 0.745 |
| **VIC** | **0.742** | 0.515 |
| **VICG** | n.a. | 0.463 |
| **OBJ** | 0.624 | **0.78** |
| **LOC** | 0.591 | **0.683** |
| **PAR** | 0.84 | **0.928** |

## Experiments on Extractive QA

- **Gold Standard:** 30 manually annotated news articles.

| Method | Model | Exact Match | | | Partial Match | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| *Single-Span QA* | *BERT$_{BASE}$* | 23% | 13% | 17% | 39% | 34% | 36% |
| | *UmBERTo* | **27%** | **15%** | **20%** | 40% | 40% | 40% |
| | *ELECTRA* | 27% | 15% | 20% | **41%** | **40%** | **41%** |
| *Multi-Span QA* | *BERT$_{BASE}$* | **36%** | **21%** | **26%** | **41%** | **27%** | **32%** |
| | *UmBERTo* | - | - | - | - | - | - |
| | *ELECTRA* | 30% | 17% | 22% | 37% | 24% | 29% |
| Human Annotation | | 88% | 84% | 86% | 91% | 89% | 90% |

## Conclusion and future work

DICE is a new corpus containing more than 10k crime news with automatic annotations and a 1000 documents manually annotated according to a **new annotation schema** proposed to extract the main entities involved in a crime event.

Future work:

- Extend the annotation schema for **new crime categories** (murder, aggression, robbery,…).
- To Increase the annotated docs through manual annotation, projection of annotation and the creation of a synthetic dataset.
- Usage of the datasets in different tasks: Multi-Span QA, Text Categorization and others.

## References

Maria Pia di Buono, Martin Tutek, Jan Šnajder, Goran Glavaš, Bojana Dalbelo Bašić, and Nataša Milić-Frayling. 2017. Two Layers of Annotation for Representing Event Mentions in News Stories. In Proceedings of the 11th Linguistic Annotation Workshop. Association for Computational Linguistics, Valencia, Spain, 82–90. https://doi.org/10.18653/v1/W17-0810

Rollo, Federica, and Po, Laura. (2020) Crime Event Localization and Deduplication. In: Pan J.Z. et al. (eds) The Semantic Web – ISWC 2020. ISWC 2020. Lecture Notes in Computer Science, vol 12507. Springer, Cham. https://doi.org/10.1007/978-3-030-62466-8_23

Rollo, F., Bonisoli, G., Po, L. (2022). Supervised and Unsupervised Categorization of an Imbalanced Italian Crime News Dataset. In: Ziemba, E., Chmielarz, W. (eds) Information Technology for Management: Business and Social Issues. FedCSIS-AIST ISM 2021 2021. Lecture Notes in Business Information Processing, vol 442. Springer, Cham. https://doi.org/10.1007/978-3-030-98997-2_6