Crime event localization and deduplication

Federica Rollo[0000−0002−3834−3629] and Laura Po[0000−0002−3345−176X]

'Enzo Ferrari' Engineering Department
Via Vivarelli, 10, 42125 Modena, ITALY
name.surname@unimore.it

Abstract. Crime analysis is an approach for identifying patterns and trends in crime events, while information extraction is the task of extracting relevant information from unstructured data. If crime reports are not directly available to the public, a possible solution is to derive crime information published in newspaper articles.
This paper aims at extracting, localizing, deduplicating, and visualizing crime events from online news articles. This work demonstrates how crime-related information can be obtained from newspapers and exploited to create a consistent database of crime events with an automatic process. The approach employs a Named Entity Recognition (NER) algorithm to retrieve locations, organizations and persons and a mapping phase to link entities to Linked Data resources. The date of the event is retrieved through the temporal expressions extraction and normalization. For duplicate detection, an approach analyses and combines crime category, description, location, and crime event date to identify which news articles refer to the same event. The approach has been successfully applied in the Modena province (Italy), focusing on eleven types of crime happen from 2011 till now. The flexibility of the approach allows it to be easily adapted to other cities, regions, or countries and also to other domains.

Keywords: Crime analysis · Crime mapping · News extraction · Document similarity · Duplicate detection.

1 Introduction

Focusing resources on high-crime places, high-rate offenders, and repeated victims can help police effectively reducing the crime rate in their communities. Police can take advantage of knowing when, where, and how to focus its resources, as well as how to evaluate the effectiveness of their strategies. Sound crime analysis is paramount to this success. Crime analysis is not merely crime events counting; it is an in-depth examination of the different criminogenic factors (e.g., time, place, socio-demographics) that help understanding why the crime occurs. Data-driven policing and associated crime analysis are still dawning. The use of Geographic Information System (GIS) techniques helps crime analysis and allows localizing crimes to identify the high-risk areas. Unfortunately, in some countries (e.g. in Italy), authorities do not provide free access to

updated datasets containing information about crimes happening in the cities. Extracting crime events from news articles published on the web by local newspapers can help overcome the lack of crime up-to-date information.

Several countries provide statistics on crime, but they are often available with some delay. In most of the cases, they are provided as aggregated data, not as single crime events. In the UK, open data about crime and policing are available at street-level with a delay of 2 months[1]. In Italy, the reports of the Italian National Institute of Statistics (ISTAT)[2] provide a clear picture of the types of crime happen in each province during the year. The information provided is aggregated by time and space and become available after (at least) one year from the crime event happening. Based on these official statistics, it is not possible to perform an up-to-date analysis of the local situation in each neighborhood. Newspapers instead provide reliable, localized, and timely information (the time delay between the occurrence of the event and the publication of the news does not exceed 24/48 hours). The main drawback is that newspapers do not collect and publish all the facts related to crimes, but only the most relevant, i.e., the ones that arouse the interest of the readers. Therefore, there is a percentage of police reports that will not be turned into news and is lost.

This paper presents a data ingestion approach for extracting crime data from news and enriching them with semantic information. The strategy employs several techniques: crime categorization, named entity extraction, linked data mapping, geolocalization, time expression normalization, and de-duplication. From the best of our knowledge, the integration of multiple techniques, previously used in different contexts, for solving various sub-problems into a common framework in order to perform crime analysis is a novelty.

The method has been applied and tested in the Modena province, a 2,688 Km^2 area populated by more than 700 000 inhabitants and located in the Emilia-Romagna region in Italy. On 13000 reports, collected from 2011 to now (May 2020), the approach was able to geolocalize almost 100% of the crime events and normalize the time expressions on 83% of the news articles. The results produced allows performing crime mapping studies and the identification of crime hot spots in semi real-time. Some preliminary visualizations of these results are shown through a web application.

The paper is organized as follows: Section 2 introduces related work; Section 3 describes the general approach to extract and analyze news articles; in Section 4 the method implemented in the Modena province is described in detail and then evaluated in Section 5. Section 6 shows the effectiveness of our approach and demonstrates its scalability. In the end, in Section 7, conclusions and possible future work are sketched.

[1] On the police open data portal https://data.police.uk/, it was possible to download data about March 2020 on the 21st May 2020.

[2] https://www.istat.it/en/

2 Related Work

In the last decades, methods for news content extraction have gained relevant interest [20, 4], and several online platforms have been developed to visualize the results of the extraction, such as the Europe Media Monitor (EMM) NewsExplorer[3] and NewsBrief[4], the Thomson Reuters Open Calais[5], and the Event Registry[6]. These platforms download news articles from the web and exploit different language processing algorithms to detect entities, group news articles into clusters according to their topics [7, 12]. In particular, a lot of scientific research is devoted to crime data mining, and new software applications have been created for detecting and analyzing crime data. In [8], an approach is described to extract important entities from police narrative reports written in plain text by using a SOM (self-organizing map) clustering method. Crime analysis methods are applied to find trends [19] or to predict crime events [9], by using neural networks, Bayesian networks, and algorithms as K-nearest-neighbour, boosted decision tree, K-means. An interesting example of crime analysis and mapping in the city of Chicago[7] is explained in [3]. In this case, crime data are extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system, composed of relational databases that allow law enforcement officials to cross-reference available information in investigations and to analyze crime patterns using a geographic information system (GIS). To protect the privacy of crime victims, addresses are shown at the block level only, and specific locations are not identified.

In Italy, two cities provide updated crime datasets on their open data portals. Torino AperTO Open Data Portal includes information about the type of crime, the location and the date of the crime events occurred, but with a delay of two years[8], while Trento provides the annual burglary rate of the previous year[9]. In both cases, data are not timely, and, in Trento, they are also aggregated. In Italy, if it is not possible to get direct access to police reports, the only timely sources of crime data are newspapers that are freely available online for everyone. Our approach aims at extracting crime events from news articles, structuring the information, geolocating them, and linking them to Linked Data resources. The collected data are then published online in a web application to provide a real-time overview and some analysis of the crime situation in the Modena province.

3 The Approach

The approach that we have selected to extract semantic information starting from the news articles published on the web consists of 7 phases. It is a general

[3] https://emm.newsexplorer.eu/
[4] https://emm.newsbrief.eu/
[5] https://www.crunchbase.com/organization/opencalais
[6] https://eventregistry.org/
[7] https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6
[8] http://aperto.comune.torino.it/
[9] European Data Portal https://www.europeandataportal.eu/it

approach that can be applied to any information content that describes events. The phases should be executed in sequence for each news, but different news articles can be processed in parallel.

The first phase is the data extraction in which information of interest that is published on the web is harvested. Then this content is labeled and structured to be stored in a database and to be semantically annotated. Some web content may already expose a structure. For example, HTML pages encoded with the Document Object Model (DOM), have a tree structure wherein each node is an object representing a part of the document.

The second phase is the Named Entity Extraction. This phase is crucial since it allows us to identify persons, organizations, places, and temporal expressions in the text of the news. The correct identification and extraction of entities are of great importance not only for enriching the crime description but also because identified entities act as annotations for the crime event.

The third phase is the Linked Data Mapping. This phase maps the entities into Linked Datasets. In particular, persons, organizations, and locations are linked to URI.

The fourth phase is the categorization of the criminal event. This phase is crucial to map a news w.r.t. a type of crime. Given some pre-categorized news, i.e., annotated training data, machine learning algorithms can be applied on uncategorized news to assign them a type of crime. The entities extracted in the previous two phases can be exploited to enhance the results of this phase.

The fifth phase is the geographical localization; in this phase, the entities that have been identified as locations are processed to be georeferenced. Different methods can be applied; moreover, if a location is not specified in the news, organizations can also be exploited to geolocate the event.

The sixth phase is the normalization of temporal expressions. In this phase, the date of the news published on the web can be revised to identify the exact time of the crime. By analyzing the news text, temporal expressions can be identified that allow normalizing the date (for example, words like “yesterday”, “two days ago”, “this morning” are identified as temporal expressions).

The last phase is the identification of duplicates. This phase should be applied not only in the case that the input sources are more than one since it is possible to find the same event described in more news articles also within the same newspaper where updates about one crime event are published along time. The duplicate detection phase is carried out by identifying possible duplicates and then making a comparison on the news text to confirm that they are duplicates. At the end of this phase, after identifying the duplicates, it is also possible to merge the information of the criminal events to enrich the information that will be stored on the database.

The use of semantic technologies is a key point in the presented approach for adding knowledge about the crime events, geolocalizing crimes, and performing deduplication. The NER combined with Linked Data mapping allows retrieving entities and associating them to stable URIs. Besides, also the deduplication takes advantage of the semantic information extracted in the previous phases.

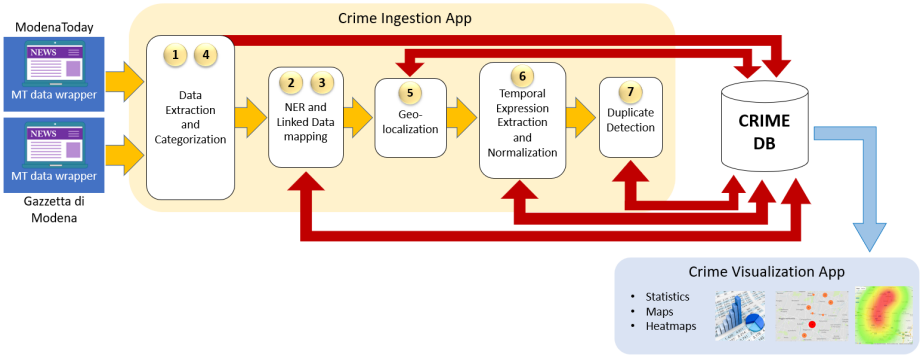

Fig. (1) The method implemented to extract, store, analyze, and visualize the news articles for Modena province (the numbers in the circles represent the phases described in Section 3).

4 Modena Crime Ingestion

The approach described in Section 3 has been implemented to extract crime-related information in the Modena province. We extended our previous work [13, 15] that collected and analyzed the news articles extracted from one newspaper related to thefts. Currently, we ingest news coming from the two most popular local newspapers, "ModenaToday" (MT) and "Gazzetta di Modena" (GM), related to eleven types of crimes: theft, attack, drug dealing, evasion, fraud, abuse, murder, robbery, money laundering, kidnapping, and sexual violence. More than 13000 reports have been collected from 2011 to now (May 2020) using this approach. There exist other 3 minor online newspapers, however integrating them will not change substantially the results since they cover around the 5% of the total news. Since the scope of the news published in these newspapers is to provide information related to single events, a single-crime-event-per-document assumption is made.

Figure 1 displays the phases of our approach applied in the Modena scenario; some phases are performed together because of the particular structure of the data taken into consideration. The Crime Ingestion App aims at extracting, parsing and storing information of the news articles into a PostgreSQL database, called CRIME DB, whose structure is discussed in detail in the following; while, the Crime Visualization App displays the information stored in the DB trough interactive crime maps, heatmaps to discover the high-crime areas, and statistics to identify trends.

The Crime Ingestion App has been implemented through a Java application by using suited libraries and APIs, while, the Crime visualization App is a web application implemented in Python[10].

[10] The code of both applications is open source and is available in a github repository https://github.com/federicarollo/Crime-event-localization-and-deduplication.

4.1 Data extraction and Categorization

The first phase of the Crime Ingestion App is data extraction. This phase is in charge of crawling the web page content to extract semantic information such as the publication date, the location, and a textual description of the event.

In our case, Modena newspapers already classify news articles according to the crime type. Thus, the categorization of news is performed within the data extraction phase. Data extraction takes in input the url of the web page containing a list of news articles related to a specific type of crime[11], then automatically retrieves the URLs related to each news, accesses each URL and extracts information from the HTML tags by using the java web crawler Jsoup.

We extract information from the news exploiting the Document Object Model and taking advantage of the HTML tags of the web page. In the Modena newspapers taken into consideration, the type of crime, the title, the subtitle (description), the date and time of the publication of the news, and the textual information can be harvested directly from the newspapers with specific HTML tags. In some cases, also the location is reported; this information is usually the name of the city or may contain the area and the address. All this information is stored in the "news" table and then refines in the following phases. The structure of the CRIME DB used to store crime-related news is displayed in Figure 2. The database is a PostgreSQL database that uses the extension PostGIS to store geospatial data. In the "news" table all the information extracted from the news is stored.

4.2 NER and Linked Data mapping

We implemented the NER phase by using Tint[12] (The Italian NLP Tool) [10], an open-source tool for NLP of Italian texts. Tint is a collection of modules customized for the analysis of text in Italian language and based on the Stanford CoreNLP. In [10] Tint was compared with three other NLP tools for Italian language, (Tanl [5], TextPro [11] and TreeTagger [16]) and it reported the higher values of speed, precision, and F-measure for the NER task (in particular, recognition of entities including persons, organizations, and locations, which is our scope). The NER module of Tint uses the Conditional Random Field (CRF) sequence tagger included in the Stanford CoreNLP, and it is trained on the ICAB dataset, which contains 180K words taken from the Italian newspaper "L'Adige" [10].

We used Tint to detect persons, organizations, locations, and time expressions. Each entity recognized by the NER algorithm is stored as an instance of the "entity" table and linked to the news in the "news" table (see Figure 2). For the entities retrieved, we perform a mapping w.r.t. linked data, in particular DBpedia and Linked Geo Data. For each person, organization and location retrieved, we search for a corresponding entity in DBpedia. An http request is sent

[11] An example is available at https://gazzettadimodena.gelocal.it/ricerca?query=furti where *furti* (theft) is a type of crime.

[12] http://tint.fbk.eu/

to the link obtained concatenating the base URI of DBpedia[13] to the name of the entity formatted with mixed case with underscores. If the request is successful, the URI is stored in the "entity" table. For each location and organization, we search for the corresponding resource in Linked Geo Data [17]: a SPARQL query selects all the spatial resources that are located in the area of Modena province and looks for the best matches. If a municipality is defined for the news, the query performed retrieves the Linked Geo Data resources of that municipality.

4.3 Geolocalization

The GPS coordinates of the crime locations are retrieved by using the OpenStreetMap API, and in particular, Nominatim[14]. Geolocation is an iterative process; it starts by evaluating whether municipality and addresses are populated in the DB and trying to geolocalize them. If the municipality or address is not present or if their geolocalization failed[15], the process starts exploring the entities that have been extracted with NER and are stored in the entity table. When multiple locations are detected, we consider the one with higher frequency. If the geolocalization of locations failed, organizations are explored. If more locations/organizations with the same frequency are available, we take the first occurrence in the news. This does not affect significantly the result of the geolocalization; indeed, we checked manually some cases, and we discovered that, in the majority of the cases, locations in a news were all close together. In the end, the GPS coordinates and the related address are stored in the "news" table.

4.4 Temporal Expression Extraction and Normalization

To detect when the crime event described in the news happen, an algorithm of temporal expression extraction and normalization is applied to the concatenation of title, sub-title, and text attributes. We use the HeidelTime temporal tagger [18] and, in particular, its implementation included in Tint for the news document type. This tool can extract temporal expressions from documents in natural language and normalize them according to the TIMEX3 annotation standard. Using the date of publication of the news ("publication_datetime") and the result of the temporal expression normalization, the date of the crime event is calculated and stored in the "event_datetime" attribute in the "news" table. When multiple event dates are detected, we select the one with higher frequency. In case of multiple event dates with the same frequency, we consider the date closest to the publication date.

[13] http://dbpedia.org/resource/

[14] https://nominatim.openstreetmap.org/

[15] We use the function *OpenStreetMapUtils.getInstance().getCoordinates(location)* where *location* is a string that can be generated by the municipality and the address, or the entity name retrieved from the DB. This function provides the latitude and the longitude of the location. The success of this function depends on how the address is stored in Open Street Map and how the location is reported in the news.

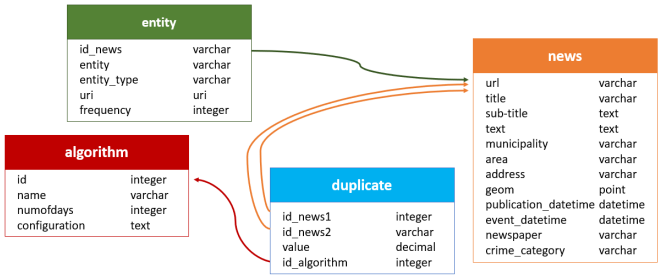

Fig. (2) The structure of the CRIME DB.

4.5 Duplicate Detection

Considering two or more newspapers of the same geographic area, there is a high probability that the news articles describing the same event are published in both newspapers the same day or a few days later. Besides, the same crime-related event could be described more times in the same newspaper to provide updates. Duplicate detection (also known as *deduplication*) is a fundamental task when we are ingesting news articles. All the news articles identified as *duplicates* are collected: in the "duplicate" table, the links to the identifiers of the two news articles, the value of their similarity and the identifier of the algorithm used to find them are stored. In the "algorithm" table, instead, the features and the parameters used by the duplicate detection algorithm are stored in the "configuration" attribute. The duplicate detection is performed in two steps: (1) the reduction of the number of news to be compared, and (2) the comparison of the news articles through similarity algorithms. After the identification of *duplicates*, a merge of the information of the two news is made.

Reduction of the number of pairwise comparisons This task is also known with the name of blocking. The identification of the event date allows us to compare only the news with the same event date to search for duplicates. If no event date is available, the comparisons are performed by using a date slot which considers the publication date. We assume that news articles related to the same crime event might be published with a few days of delay in different newspapers or the same newspaper. Therefore, we do not need to compare a huge amount of news articles. Besides, the blocking technique we applied compares only news articles classified as the same type of crime and related to crimes that happened in the same municipality, including cases in which municipality is not specified.

Comparison of news articles through similarity algorithms The similarity between two news articles takes into account the semantic information extracted in the previous phases and, in particular, the municipality and the event date, in addition to the title, sub-title, and text. In particular, the simi-

larity is computed by using the following multi-variables formula:

$$similarity = \alpha * X + \beta * Y + \gamma * Z + A + B \quad (1)$$

where X is the similarity between titles of the two news articles, Y is the similarity between sub-titles, Z is the similarity between texts, and α, β, γ are configuration parameters representing the weights to be assigned according to the importance of each similarity; in the end, A is a value added only if the two news articles have the same municipality and B if they have the same event date or the same publication date if the event date was not extracted. The resulted value is normalized and compared to the threshold T. The latter is a configuration parameter that determines if the two compared news articles are related to the same crime event or not. If the similarity computed by the multi-variables formula is equal or greater than T the news articles are labeled as duplicates, otherwise they are considered as related to different events.

For duplicate detection, we apply the shingling technique [6]. A shingle is a portion of the text, also known as "n-gram", where n is the number of consecutive words in the text. In the sentence *"The mugger escaped with the bag"*, if we use the 3-Shingling technique, the shingles are "the mugger escaped", "mugger escaped with", "escaped with the", and "with the bag". We applied the Jaccard index and the cosine similarity methods [1] on a set of 50 news articles manually labeled, and we analyzed the obtained results. We concluded that the cosine similarity computed by using the shingling technique is the best option since it provides better results of recall, precision, and accuracy. This result could be expected since the Jaccard index is not affected by the number of word occurrences. The cosine similarity instead takes into account how many times the shingle appears in specific news and its "importance" thanks to the TF-IDF (Term Frequency - Inverse Document Frequency) function. The "importance" of shingle is made by how many times it appears in the news (i.e., *term frequency*), compared to how many times it appears in all the collection of news articles to be compared (i.e., *document frequency*). The duplicate detection exports data from the DB, applies the similarity algorithm, stores the results, and merges the duplicates. We use the library "java-string-similarity"[16], a Java library implementing different string similarity and distance measures, including shingle based algorithms with cosine similarity.

To determine the best values for the configuration parameters, a validation test has been made on a dataset containing 358 news articles related to robberies published between February 2019 and May 2019 on both newspapers (in particular, 196 news articles are from "ModenaToday" and 162 news articles from "Gazzetta di Modena"). We have read and analyzed the news articles manually and have found 44 duplicates. Then, we have listed the identifiers of these duplicates to be compared with the results of our algorithm.

Tests have been performed between news articles with the same event date, and also considering news articles without event date published at most three or five days before (respectively, *3-days slot* or *5-days slot*). The news articles pairs

[16] https://github.com/tdebatty/java-string-similarity

Table (1) Validation test of the *3-days slot* duplicate detection

T	Accuracy	Precision	Recall	F1-score
0.71	0.92	1.00	0.36	0.53
0.70	0.94	0.79	0.68	0.73
0.69	0.92	0.68	0.68	0.68
0.68	0.91	0.63	0.70	0.67
0.67	0.91	0.60	0.82	0.68
0.66	0.87	0.49	0.84	0.62
0.65	0.86	0.47	0.98	0.64

Table (2) Validation test of the *5-days slot* duplicate detection

T	Accuracy	Precision	Recall	F1-score
0.71	0.91	0.89	0.37	0.52
0.70	0.93	0.72	0.70	0.71
0.69	0.92	0.66	0.70	0.68
0.68	0.91	0.61	0.72	0.67
0.67	0.90	0.57	0.84	0.68
0.66	0.86	0.46	0.86	0.60
0.65	0.84	0.43	1.00	0.60

to be compared selected by the reduction algorithm are 376. At this point, we have to choose the dimension of the shingles. Obviously, the smaller the shingle size, the greater the amount of storage capacity and space complexity required. In [2], Alonso et al. discovered that the optimal shingle dimension without losing precision was between 1 and 3. Based on these tests, we set the length of the shingle to 2 for the title and the sub-title and 3 for the body of the news. Then, we tried to determine the weights α, β, γ, and the values of A and B. After tests with different values, we chose the values $\alpha = 1$, $\beta = 1.25$, $\gamma = 4$, $A = 0.025$, and $B = 0.03$ since they provided the higher values of precision, recall, and accuracy. Therefore, we selected these as the best values. Table 1 shows the evaluation of tests using the slot of 3-days with different values of threshold T. Experiments described in Table 2, instead, are made using the slot of 5-days. In both cases, the best results are obtained with the threshold set to 0.70. The differences in accuracy, precision, recall, and F1-measure are minimal. The time required by the duplicate detection algorithm is a little more than one minute.

5 Evaluation

To evaluate the efficiency of the approach, we performed three tests: one test to check the efficiency and effectiveness of the NER algorithm and the use of Linked Geo Data for geolocalization, another one to evaluate the temporal expression extraction and normalization phase and the last test to access the deduplication algorithm. The data extraction phase implements a high precision method and always extracts the information of interest (title, sub-title, text, location, date and time of publication, type of crime) in the right way since the approach is based on the HTML tags of the source of the news.

5.1 NER effectiveness

The NER was evaluated on 530 news articles manually labeled, published on both "ModenaToday" and "Gazzetta di Modena" newspapers from January 2020 to March 2020 related to all the types of crime taken into consideration. The manual annotation of the news identifies only one location for each news article.

To prove that the adoption of the NER and the mapping w.r.t. Linked Geo Data enhance the geolocalization, we check the data we obtained without the application of these methodologies. Without NER, we can only make use of the tags on the HTML documents. In a newspaper like "ModenaToday", only the city or the area within the city to which the news refers are provided. In this case, we can establish a correspondence with the municipality for 95% of news articles. This reference is very loose because it does not allow us to locate the crime event to a specific point but only to refer it to an area/city. On the other hand, by using NER, we can extract persons, organizations and locations from each news article. The geolocalization of the location entities allows us to retrieve the GPS coordinates in 479 news (90%). If no location is discovered, the geolocalization is applied to the organizations, finding 34 geolocated entities. Searching in Linked Geo Data the locations/organizations, a location for each news is found. Moreover, with Linked Geo Data, each crime event is enriched and linked to a URI referencing its location. Table 3 shows how the application of NER and the integration with Linked Geo Data increase the possibility of retrieving the GPS coordinates. In the table, the first column explains the method used, the second column is the number of news articles where the location reference is found, the third column shows how many of these locations are geolocated in the area of Modena province, and in the end, the last column is the number of news articles where a location reference is not found. As can be seen, Linked Geo Data allows finding a location reference for all news in the dataset and revises the coordinates geolocated out from the area of Modena province. Without the use of Linked Geo Data, the locations and organizations, identified with the NER, allow finding location reference in 96% of the news articles in the dataset.

Table (3) Evaluation of the coverage in detecting location reference through the application of NER and the integration of Linked Geo Data

method	location references found	references localized in Modena province	location references not found
NER locations	474	454	55
+ NER organizations	+37 (511)	+46 (500)	-37 (18)
+ Linked Geo Data	+18 (529)	+29 (529)	-18 (0)

All the locations retrieved are geolocated in Modena province thanks to the integration with Linked Geo Data. Due to the incompleteness of some addresses found by the NER locations and organizations, the OpenStreetMap API is not able to retrieve the GPS coordinates. In some cases, the crime events are geolocated out from the Modena province. Linked Geo Data revises these errors. Comparing the results with our manual annotations, the geolocalization performs with a 90% precision.

Considering all the news in our dataset discovered by our approach, 12482 news articles among the 13751 news articles have been geolocated exploiting

the locations found by NER, 515 of these GPS coordinates were out of the Modena province. On the remaining news, 13153 news articles have been located through organizations (293 out of the Modena province) and 13725 thanks to the integration with Linked Geo Data (1 out of the Modena province). In conclusion, almost 100% of news articles have been geolocated in the province of Modena.

5.2 Temporal Expression Normalization effectiveness

The use of the HeidelTime tagger allowed extracting temporal expressions in 11426 news articles (83% of the total number of stored news articles). To evaluate the precision of such an approach, a manual evaluation was performed on the same dataset used for the evaluation of the NER effectiveness. One event date has been manually identified for each news. The algorithm has extracted and normalized time reference in 440 news articles (83%). All the time references found are correctly normalized. In the remaining 90 news articles, the identification of time reference was a bit challenging since it was embedded in expressions like "after six months", or "at dawn", or other similar expressions. In other cases, the news is about some updates of an event very far in time; therefore, only the reference to the year is found.

5.3 Duplicate detection effectiveness

The duplicate detection algorithm was evaluated on 470 news articles coming from the two different newspapers, and related to different types of crime (165 thefts, 127 robberies, and 178 attacks). The dataset was manually labeled, and 49 duplicates were found. The reduction algorithm selected 261 news articles pairs to be compared (43 about robberies, 100 attacks, and 118 thefts); the time required to find duplicates was two seconds. The deduplication has been applied using three alternatives: 3-days and 5-days time slots on the news date, and the event date. Using the time slots performed similarly with the threshold of 0.70. The results are shown in Table 4. The precision is a bit higher for the 3-days slot, where it reaches the 91%, while the recall is better for the 5-days slot, where it is 88%. The third evaluation was done exploiting the results of the Temporal Expression Normalization phase. On the test set, we found 440 news articles (94%) with a time expression that has been extracted, and normalized, and used to define the event date. The duplicate detection considering the event date in the pairwise comparison allowed discovering 95% of the duplicates. Thus, improving the results of the duplicate detection algorithm.

Table (4) Duplicate detection algorithm on different *day slots*

T	*day slot*	Accuracy	Precision	Recall	F1-measure
0.70	3-days	0.96	0.91	0.65	0.76
0.70	5-days	0.97	0.84	0.88	0.86

6 Impact and scalability of this research

In Section 4, we took into consideration two local newspapers to ingest crime events. However, newspapers do not publish news for each crime happened in Modena since some crimes are not of public interest. To evaluate the impact of our method, we made a comparison between the number of crimes in the CRIME DB and the number of crimes published in the official report of ISTAT (i.e. the crimes reported to the police). The latest report available[17] contains the crime rates per province from 2014 till 2018. The information is only quantitative; the types of crime are reported per province as a rate out of 100000 inhabitants. No information about where the crime happened (the municipality, the district, or the address), and the period of the year (season, month, or date) is provided. The classification of crimes includes 50 types of crimes and is very exhaustive, including categories that are not taken into consideration in our approach. For providing a comparison between the two datasets, we consider only the crime categories in common. The number of crimes in our approach is considered after the deduplication. Since a location-based comparison was not possible because ISTAT provides a unique report for the entire province, we calculated the number of crimes for the city of Modena in each year by using the total population registered in the same year (from 184700 till 186300 inhabitants). With the total number of crimes in the city of Modena of 7107 from 2014 till 2018, our approach covers around 21% of the crimes reported by ISTAT. A hypothesis on this low coverage can be attributed to the fact that not all the criminal events recorded by ISTAT, and therefore in the police reports, are of high impact. Therefore, not all of them are reported in local news articles.

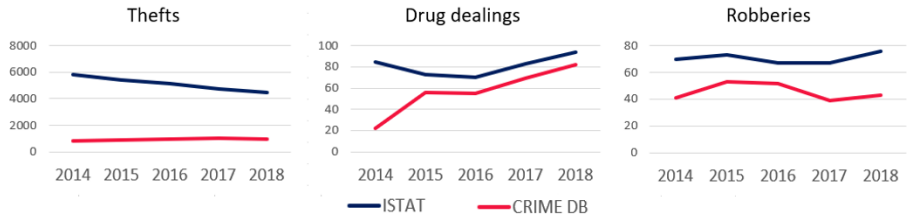

Fig. (3) Distribution of crime reports from 2011 to 2019 in different neighborhoods of Modena.

The most frequent crimes in both datasets are thefts for each year from 2014 to 2018. Figure 3 shows the total number of the top three types of crimes recorded in the report of ISTAT and compared to the number collected with our approach. As can be seen, the lower coverage is reported in thefts (the percentage average is 18%), while the higher one is in drug dealings (70%). It is very interesting to see

[17] http://dati.istat.it/Index.aspx?QueryId=25097&lang=en

that the trend over time of drug dealing is very similar in the two distributions if we do not consider the year 2014. Robberies have quite a similar trend in both the datasets (we were able to find 64% of the total robberies).

The approach here described is currently in use to identify the crimes occurring in Modena and its province in real-time. The visualization of crime-related data is available online[18]. This application is a Python web application which uses Tornado as a web framework. It is still an in-progress application since we want to integrate heatmaps to detect hot spots and some statistics. Everybody wants to avoid finding ourselves involved in unpleasant incidents. The citizens and the visitors of a city need to be aware of crime statistics; this may affect where they will stay, go for a walk, live, and work. On the other hand, city councils have the responsibility of identifying crime hot spots to employ appropriate monitoring and controlling mechanisms. Regarding the impact of this approach, even if the approach has been applied in a medium area, it highlights its potentiality. As reported in Section 1, in Italy it is not possible to collect real-time crime information from official sources, since official criminal statistics are reported annually with a delay of 6 months. The approach we have implemented can be applied everywhere also in small or medium cities/areas since there will be always one or more newspapers that report the main crimes to happen in that place.

A first scalability test has been executed to ingest all the news articles related to crimes happened in the entire Emilia-Romagna region. We selected other 9 newspapers which publish news related to the 9 provinces of the Emilia-Romagna region. We collected all the available news articles, from 2011 till now, which refer to the eleven crime types. The total number of news articles is 35000 (on average 3900 news articles for each province). The crime ingestion can be run in parallel for different newspapers and for different crime type. Therefore, we executed 99 ingestion processes in parallel to extract, analyze and store data of the region. The total loading time, that depends on the loading time of the province with the higher number of news from 2011, is 3 hours for 35000 news[19].

7 Conclusion and future work

In this paper, we integrated multiple techniques for solving various sub-problems into a common framework in order to perform crime ingestion that is the first step to allow further analysis and visualization related to crime events. Our approach aims at extracting crime events from news articles, structuring the information, geolocating them, and linking them to Linked Data resources. The collected data are deduplicated and published online in a web application to provide a semi real-time overview and some analysis of the crime situation in the Modena province.

The paper described a general procedure that consists of 7 phases and its application in the context of Italian crime news for the Modena province. The

[18] Crime Visualization App - https://dbgroup.ing.unimore.it/crimemap

[19] The test has been performed on a Microsoft Windows 10 Pro with 16GB RAM.

approach has made it possible to create a consistent dataset with more than 13000 news articles concerning 11 types of crime, to identify crime events date and location, so allowing a time-space analysis unveiling this critical data to the citizen. A comparison with the official data gave way to discover that this approach allows collecting about 20% of the crime events available in the official reports. The use of the NER algorithm combined with the Linked Data mapping has enhanced the semantic information of each crime and the geolocalization of the events. The time expression normalization has improved the performance of the duplicate detection algorithm. From the best of our knowledge, this is the first case in which a citizen of a medium Italian city can have a look at the real-time crime data and recent statistics on crime trends. The approach is domain-independent. In this paper it is applied on the context of crime-related articles; however, it is possible to apply it to any kind of news. For new sources, an individual wrapper/extractor has to be built to retrieve title, subtitle, and so on. This is the only part that needs to be built (and in some cases only adapted from other wrappers) in order to connect a new source with the application. All the other phases can be re-used as they are provided in the open-source code. The NER algorithm can be applied to new sources provided that the text is in Italian, while the temporal expression extraction and normalization can also be performed on text of other languages. The geolocation process can geolocate addresses all over the world. The duplicate detection can be applied to text in different languages since the similarity measure is not affected by the language. In the end, the Crime Visualization App can show different types of events stored in a database with the structure of the CRIME DB.

In the near future, we will add in the process *Keyphrase/Keyword Extraction* that extracts the main phrases that categorize the text [14]. Moreover, an additional phase will be integrated to detect the key elements of each news. News articles related to the same type of crime have common characteristics. For example, all news articles related to thefts refer to a stolen object (car, money, and other objects); while in the news related to the attacks there is always the mention to who was attacked, and sometimes the reference to who was the attacker and the reason for the attack. This information is specified in a phrase that is characteristic of each type of crime.

References

1. Agarwal, N., Rawat, M., Maheshwari, V.: Comparative analysis of jaccard coefficient and cosine similarity for web document similarity measure. International Journal for Advance Research in Engineering and Technology **2**(X), 18–21 (2014)
2. Alonso, O., Fetterly, D., Manasse, M.: Duplicate news story detection revisited. In: Banchs, R.E., Silvestri, F., Liu, T.Y., Zhang, M., Gao, S., Lang, J. (eds.) Information Retrieval Technology. pp. 203–214. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
3. Alqahtani, A., Garima, A., Alaiad, A.: Crime analysis in chicago city. In: International Conference on Information and Communication Systems (ICICS). pp. 166–172 (2019). https://doi.org/10.1109/IACS.2019.8809142

4. Arya, C., Dwivedi, S.K.: Content extraction from news web pages using tag tree. Int. J. Auton. Comp. **3**(1), 34–51 (2018). https://doi.org/10.1504/IJAC.2018.10013755
5. Attardi, G., Dei Rossi, S., Simi, M.: The tanl pipeline. In: Proc. of the Workshop on Web Services and Processing Pipelines in HLT, co-located LREC (2010)
6. Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G.: Syntactic clustering of the web. Computer networks and ISDN systems **29**(8-13), 1157–1166 (1997)
7. Chaulagain, B., Shakya, A., Bhatt, B., Newar, D.K.P., Panday, S.P., Pandey, R.K.: Casualty information extraction and analysis from news. In: Proc. of the International Conference on Information Systems for Crisis Response and Management. ISCRAM Association (2019)
8. Keyvanpour, M.R., Javideh, M., Ebrahimi, M.R.: Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science **3**, 872 – 880 (2011)
9. Oatley, G., Zeleznikow, J., Ewart, B.: Matching and predicting crimes. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence. pp. 19–32. Springer (2004)
10. Palmero Aprosio, A., Moretti, G.: Italy goes to Stanford: a collection of CoreNLP modules for Italian. ArXiv e-prints (2016)
11. Pianta, E., Girardi, C., Zanoli, R., Kessler, F.B.: The textpro tool suite. In: Proc. of LREC-08 (2008)
12. Piskorski, J., Zavarella, V., Atkinson, M., Verile, M.: Timelines: Entity-centric event extraction from online news. In: Proc. of Text2Story - Third Workshop on Narrative Extraction From Texts. CEUR Workshop Proceedings, vol. 2593, pp. 105–114. CEUR-WS.org (2020)
13. Po, L., Rollo, F.: Building an urban theft map by analyzing newspaper crime reports. In: 13th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP Zaragoza, Spain. pp. 13–18 (2018). https://doi.org/10.1109/SMAP.2018.8501866
14. Po, L., Rollo, F., Lado, R.T.: Topic detection in multichannel italian newspapers. In: Semantic Keyword-Based Search on Structured Data Sources - COST Action IC1302 Second International KEYSTONE Conference, IKC, Cluj-Napoca, Romania. pp. 62–75 (2016). https://doi.org/10.1007/978-3-319-53640-8_6
15. Rollo, F.: A key-entity graph for clustering multichannel news: student research abstract. In: Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco. pp. 699–700 (2017). https://doi.org/10.1145/3019612.3019930
16. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing, Manchester, UK (1994)
17. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: Linkedgeodata: A core for a web of spatial open data. Semant. Web **3**, 333–354 (Oct 2012)
18. Strötgen, J., Gertz, M.: Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: Proc. of the International Workshop on Semantic Evaluation, SemEval@ACL. pp. 321–324 (2010)
19. Wang, T., Rudin, C., Wagner, D., Sevieri, R.: Learning to detect patterns of crime. In: Machine Learning and Knowledge Discovery in Databases. pp. 515–530 (2013)
20. Zhang, K., Zhang, C., Chen, X., Tan, J.: Automatic web news extraction based on DS theory considering content topics. In: Proc. of International Conference. Lecture Notes in Computer Science, vol. 10860, pp. 194–207. Springer (2018). https://doi.org/10.1007/978-3-319-93698-7_15