Forensic Science International 381 (2026) 112828

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Entity recovery in criminal investigations: Evaluating NER and investigator labels on real case texts

Johannes Pippidis Lorentzen [a,b,*], Mads Skipanes [a,c]

[a] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Teknologivegen 22, Gjøvik, 2815, Norway
[b] DNB BANK ASA, Dronning Eufemias gate 30, Oslo, 0191, Norway
[c] Kripos/The National Criminal Investigation Service, Nils Hansens vei 25, Oslo, 0667, Norway

ARTICLE INFO

Keywords:
Named entity recognition
Information extraction
Entity resolution
Text mining
Criminal investigation
Homicide cases
Forensic text analysis
Police records

ABSTRACT

We investigate how effectively automatic named entity recognition can recover entities selected by investigators in real criminal cases. Two homicide cases and one synthetic case were processed with five NER models, and extracted entities were matched to investigator labels using a lightweight entity resolution and similarity scoring method. Police trained models aligned best with investigator annotations, while general-purpose models produced larger but less relevant entity sets. Manual validation confirmed that many labels are not text recoverable, establishing a realistic upper bound on NER performance. The results indicate that extraction and investigative relevance are distinct tasks, motivating a two-stage pipeline and future work on relevance modelling and improved resolution.

1. Introduction

In major criminal investigations, collaborative efforts involving tens to hundreds of detectives across multiple forensic disciplines generate thousands of police reports containing vast amounts of unstructured text. Effective analysis of these reports requires systematic entity extraction and cross-document linking to build comprehensive case understanding. Currently, detectives perform this entity generation manually, a tedious and time-consuming process [1–4] that may result in critical information being overlooked, as shown by Skipanes et al. [5].

Recent advances in *Named Entity Recognition* (NER), which identifies entities such as persons, locations, and organisations in unstructured text [6], have enabled automation with near-human accuracy. On the CoNLL-2003 benchmark [7], transformer models achieve precision and recall above 90%. Yet, these benchmarks cover only a few generic categories, limited insight into the diverse and context-specific entity types encountered in criminal investigations [8].

A key knowledge gap concerns how NER models perform on real investigative data. Confidentiality restrictions limit access to police reports [9], leaving unclear which entity types investigators prioritise in practice. Although frameworks such as PIVCOLF (Fig. 1) [10] define ontologies for investigative entities, their alignment with practical investigative needs remains largely untested for NER applications.

Beyond entity recognition lies an equally important task *Entity Resolution*[1] (ER), which involves identifying and merging duplicate entities in large, noisy datasets [11]. Effective entity extraction in criminal investigations requires creating canonical entities that consolidate repeated mentions. Without this step, investigators cannot reliably trace relationships between entities across case materials, limiting comprehensive case analysis.

To address these gaps, we present what is, to the best of our knowledge, the first systematic evaluation of an automated NER and ER pipeline on real investigative datasets containing entities created by detectives. These manually produced entities were never intended for machine learning and therefore are often incomplete, lack explicit text-span links, and may contain incorrect or ambiguous connections. We compare automatically extracted and resolved entities with those manually created and linked by experienced investigators. In addition, we introduce a methodology for transforming investigator-generated entities into machine learning (ML)-ready datasets, bridging the gap between domain-specific investigative practices and computational requirements for NER. This approach supports the development of fine-tuned models tailored to criminal investigations rather than generic

* Corresponding author at: Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Teknologivegen 22, Gjøvik, 2815, Norway.
E-mail address: pippidis@proton.me (J.P. Lorentzen).

[1] Entity resolution can occur at several levels: within a single document, across all documents in a case, and globally across multiple cases. Closely related is the task of entity linking, which associates resolved entities with external knowledge bases. A detailed discussion of these aspects is beyond the scope of this article.

https://doi.org/10.1016/j.forsciint.2026.112828
Received 17 November 2025; Received in revised form 4 January 2026; Accepted 16 January 2026
Available online 21 January 2026
0379-0738/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

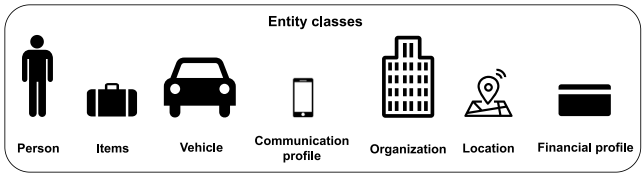


Fig. 1. PIVCOLF Ontology applied within the National Criminal Investigation Service, Norway [10].

text domains and facilitates the use of a large corpus of historical case material that would otherwise remain inaccessible for machine learning. In this study, we therefore ask the following:

> *What systematic differences exist between criminal investigators' entity selections and entities produced by an automated entity extraction system, and how can manually created investigative entities be transformed for machine learning applications in criminal investigations?*

Our contribution is twofold. First, we develop a complete and auditable pipeline that ingests investigative documents, performs NER, resolves entities at the document level, and matches the resulting entities against investigator-provided labels. Second, we validate this label-to-document matching process, identifying systematic differences between investigator annotations and automatically extracted entities. Together, these components provide a practical framework for evaluating and improving automated entity extraction in real investigative settings.

In addition, we present a systematic methodology for transforming investigator-generated data into ML-ready datasets. We demonstrate that this approach is feasible by testing it on 3848 documents and 13,630 labels drawn from two real cases. Methodologically, we argue that the NER model should identify all entities present in a document, while a separate relevance module determines which entities are of investigative importance. This architectural separation keeps entity recognition a general linguistic task, allows common NER and document-level entity resolution models to be trained on non-sensitive data, and enables secure training of a relevance module on restricted datasets.

The remainder of this paper is structured as follows. Section 2 reviews related work on named entity recognition, entity resolution, and information extraction within investigative contexts. Section 3 describes the dataset, the architecture of our pipeline, and the validation procedure. Section 4 presents the experimental results and key findings. Section 5 discusses methodological implications and outlines directions for future research, before we conclude in Section 6.

2. Related work

In this section, we review related work on named entity recognition and entity resolution within the context of criminal investigations. Both are subfields of *Information Extraction* (IE) concerned with identifying and consolidating named entities in text. Early approaches in the 1990s relied on rule-based systems and statistical models, including maximum entropy methods [12], followed by the dominance of *Conditional Random Fields* [13]. With the advent of deep learning, neural architectures rapidly outperformed traditional models. The current state of the art is defined by transformer-based architectures, introduced by Vaswani et al. [14] and exemplified by BERT [15]. These models achieve F1 scores above 90% on standard benchmarks, generating significant optimism for their application in criminal investigative contexts, as reflected in recent specialised literature [16].

2.1. NER in criminal investigations

Within criminal investigations, NER has been explored through traditional techniques such as lexical lookup and rule-based methods [17]. Most prior work has relied on open datasets with broad entity categories, while several studies have proposed domain-specific ontologies tailored to investigative contexts [18,19]. Domain-specific entity classes such as suspects, crime locations, drugs, and weapons have been introduced, and models have been trained for these categories using Norwegian data [20]. However, practitioners have noted that generic categories, including person and location, tend to be more stable and better suited for fine-tuning. From an investigative standpoint, domain-specific labels such as suspect or crime location are problematic because their meaning can shift as a case evolves: a location initially recorded as a crime scene may later be dismissed, and a person identified as a suspect may be cleared. For this reason, we focus on the generic investigative categories building upon the work of Nissen [21], and defined in the PIVCOLF ontology proposed by Skipanes et al. [10]: **P**erson, **I**tems, **V**ehicle, **C**ommunication Profiles, **O**rganizations, **L**ocation, and **F**inancial Profiles.

Beyond taxonomic design, the application of NER in criminal analysis raises fundamental linguistic and epistemological challenges. Gianola [22], based on an in-depth corpus study of 370 French witness interviews, demonstrates the heterogeneity of investigative texts, the importance of genre awareness, and the necessity of maintaining analyst control over interpretation. Complementing this perspective, [23] propose a six-stage operational framework for deploying NER in investigative environments, formalising the interaction between automated extraction, link discovery, and human validation. Our study extends these foundations through large-scale empirical evaluation, quantitatively measuring how automatic NER aligns with investigator-generated labels on real homicide case material, thereby establishing practical performance bounds for investigative NER systems.

Recent work has introduced complete NER systems for entity identification in investigative text [9,24], along with pipelines for knowledge graph construction [25]. Large language models have also been applied to related tasks, such as identifying persons and describing their roles in criminal investigations [26]. Furthermore, we [27] found that NER remains a core component of modern case management systems.

2.2. Entity resolution

Within investigative analysis, ER plays a critical role in consolidating repeated mentions into canonical entities that can be traced across documents and cases. While essential for building consistent and searchable representations of people, locations, and other objects, ER has received comparatively limited attention in the context of criminal investigations.

The most mature implementations appear in large-scale administrative systems such as the *Criminal Justice Administrative Records System* (CJARS), which applies probabilistic matching to link millions of records across heterogeneous criminal justice databases [28]. However, such systems primarily address record linkage between databases rather than resolving entities within investigative documents.

A smaller body of research has begun addressing ER for investigative text. Xing and Chen [29] proposed a two-stage alignment approach for entity extraction in police reports, while [30] introduced graph-based techniques for identity resolution across investigative materials. Robinson [31] developed methods for detecting duplicate entities in criminal networks using reference graphs to integrate information across law enforcement datasets. Together, these studies demonstrate growing interest in applying ER methods to unstructured case data, although practical implementations remain rare.

2.3. Training datasets

The development of training datasets for investigative NER has long represented a major bottleneck [9]. Creating such datasets is labor-intensive and costly, as each annotation requires domain expertise and strict quality control. To date, no studies have been identified that transform datasets originally produced for investigative purposes into datasets suitable for machine learning. Previous research has instead relied on synthetic data, often generated through large language models, and recent work has proposed taxonomies for applying LLM-based annotation methods [32]. We used an LLM for synthetic data generation and fine-tuned a domain-specific Norwegian NER model for criminal investigations [10]. However, the reliance on synthetic data highlights a persistent lack of real-world, investigator-generated datasets for model training and evaluation.

3. Methods

This section presents the methodological framework used in this study. We describe start with an illustrative example, before describing the dataset, and the evaluation pipeline. The workflow includes document ingestion, entity extraction and resolution, and systematic comparison with investigator annotations. The section begins with an illustrative use case to demonstrate the intended outcome of automated entity extraction within investigative documents.

3.1. Illustrative example

To clarify the purpose of the proposed approach, we include a short example based on a synthetic police report. The excerpt demonstrates how entities within an investigative text are identified and categorised according to the PIVCOLF ontology. The text is as follows:

> On Monday, April 26th, 2024, while serving as Detective Officer at the Stavanger police station, I, Nils Nielsen was on duty when at 3:01pm, **OLE OLSEN**, date of birth 22.02.80 visited the duty office to report a fraud. Olsen claimed he had transferred 300,000 NOK to an individual named **LARS LARSEN**, date of birth unknown, for a **BMW i3** advertised on **Facebook** marked. The vehicle was listed under the license plate **AB12345**. Olsen had engaged with Larsen through his Facebook account "**LT LARS**" and completed the payment to a **Sparebanken** account numbered **1234.33.4321** and was told to pick up the vehicle at **Stavangergata 1**. After the transaction was completed, the Facebook account "LT LARS" was deactivated. Olsen visited Stavangergata 1, and found a **note** on the wall that read 'Idiot'. This prompted him to report to the police.

Table 1 summarises the extracted entities that have been highlighted in the text, along with their corresponding case database entries. It is worth noting, however, that while only the relevant entities are highlighted, numerous additional named entities appear in the text. In real investigations, label creation also involves enrichment: investigators supplement entities with information drawn from registries, such as organisation numbers, registered addresses, or other identifying attributes, even when these details do not appear in the document text itself.

3.2. The dataset

In this study we executed the full pipeline on an air-gapped corpus of 3848 documents drawn from two real-world Norwegian homicide cases from within the last two decades. These datasets consist of various document types, including interview transcripts, evidentiality analyses, statements, and other case files. Because the dataset is highly sensitive, the exact composition of documents and time-frame cannot be disclosed. Similarly, no raw text extracts, labels, or models from the operational cases can be made available. To mitigate this, all examples in this article are based on a synthetic case, and all code needed to reproduce results on that synthetic case is provided upon request.

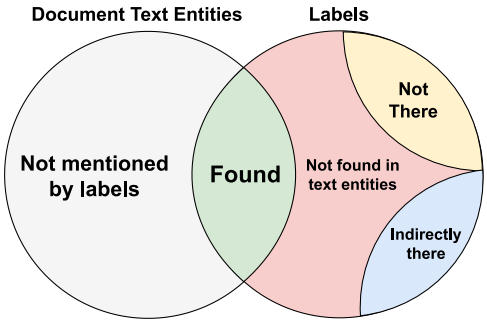


Fig. 2. Conceptual Venn diagram illustrating the relationship between entities extracted from the document text and investigator labels. The left region represents text entities. The right region represents investigator labels, including subdivisions for labels that are not present in the text and labels that are only indirectly supported by the text.

The synthetic case was created providing an LLM with the book *"The Mysterious Affair at Styles"* by [33]. The novel was used as a narrative basis to generate five case-style documents, which were then manually curated to include common forensic attributes not present in the original narrative (e.g., phone numbers). A senior investigator annotated the synthetic documents using the same labelling approach as in an operational settings.

3.2.1. The labels

At the core of the utilised dataset are the labels: the annotations, notes, and cross-references that police investigators attach to individual documents within the investigative system. These labels represent the investigators' interpretive understanding of entities, relationships, and events. They may clarify a person's identity, provide supplementary contextual information, or link related objects and occurrences across documents.

Importantly, these labels were never designed for use in machine learning. They were created for investigative reasoning and operational traceability, not for text-span alignment or computational analysis. Consequently, they lack explicit links to the exact textual spans from which they originate, and their phrasing often differs substantially from the text itself.

Because of this, there is no guarantee that a labelled entity can be directly rediscovered in the document text. For instance, a report may describe a crime scene object as *"white van"*, while the associated annotation records it as *"Ford Transit, AB12345"*. In such cases, the annotation is considered *indirectly* present: the concept is represented in the text, but the surface form does not match. Conversely, when no textual overlap or paraphrastic relationship can be identified, the annotation is categorised as *not present*. This is visualised in Fig. 2.

This distinction between direct, indirect, and not present labels is crucial for evaluating the performance of automatic entity extraction systems on real investigative data, since it defines the upper bound of achievable matching accuracy. The labelling practices of investigators thus form both the ground truth and a fundamental limitation when comparing human and machine entity recognition.

3.3. Evaluation pipeline

This section introduces the system architecture. We developed a purpose-built pipeline on real investigative cases, operating in an air-gapped environment with approved access controls. The pipeline has five major steps; Ingestion, Named Entity Recognition (NER), Entity Resolution (ER), Matching, and Validate as shown in Fig. 3. Below, we describe each component in detail.

Table 1
Example of extracted entities from a synthetic police report following the PIVCOLF ontology. The *Label in Database* column refers to how it would be labelled by the investigators in a real world case.

Entity type	Named entity	Sentence example	Label in database
Person	Ole Olsen Lars Larsen	...at 3:01pm, OLE OLSEN... ...named LARS LARSEN...	Ole Olsen, birth 22.02.80 Lars Larsen, birth unknown
Item	Note	...found a note on the...	Note: "Idiot"
Vehicle	BMW i3	...a BMW i3...plate AB12345...	BMW i3, AB12345
Communication Profile	LT LARS	...account "LT LARS" and...	LT LARS
Organisation	Facebook Sparebanken	...on Facebook marked... ...a Sparebanken account...	Facebook, Org.nr 456654 Sparebanken, Org.nr 123321
Location	Stavangergata 1	...at Stavangergata 1...	Stavangergata 1, 0989 Stavanger
Financial Profile	1234.33.4321	...account numbered 1234.33.4321	1234.33.4321



Fig. 3. High-level pipeline for document ingestion, entity extraction, matching, and validation.

3.3.1. Ingestion

Cases were first discovered by the system, and all files were registered in the database together with their associated metadata. Document and label text were extracted using the `MarkItDown`[2] Python package. All extracted text was normalised and stored in a structured format to support the subsequent NER and analysis stages.

3.3.2. Named entity extraction

The NER stage forms the first computational step in transforming raw investigative text into structured information. For every document in the corpus, the pipeline identifies span-level mentions of *Text Entities*. Each detected mention is stored together with its character offsets, entity type, source model, and a short context window to support downstream processing.[3] For comparability, text entities were also extracted from the investigator labels.

To assess the robustness of the approach across different model families and training regimes, we applied five NER models. Using multiple models allowed us to evaluate consistency, identify systematic differences, and compare the behaviour of general-purpose and domain-specific systems. The models were as follows:

ScandiNER: A general transformer-based NER model for Scandinavian languages, available via Hugging Face [35]. It represents a strong, multilingual baseline and provides broad coverage across common entity types.

Police 1 & 2: A domain-specific transformer NER model trained specifically for the PIVCOLF ontology [10]. These models reflect operational annotation practices and have been fine-tuned on representative investigative corpora.

Police 3: A domain-specific transformer NER model trained on a broader investigative ontology that extends well beyond the core PIVCOLF categories [20]. In addition to persons, locations, organisations, and identifiers, this model includes categories such as drugs, weapons, products, and event-related entities, resulting in a wider extraction scope than Police 1 and Police 2.

spaCy: A spaCy Norwegian model built on spaCy's CNN-based tok2vec architecture rather than a transformer [36,37]. Trained mainly on Norwegian news text, it provides a broad, general-purpose baseline for common entity types.

3.3.3. Entity resolution

After extracting span-level *Text Entities*, the pipeline applies a lightweight form of entity resolution to group surface variants that refer to the same real-world entity. The purpose of this step is twofold: (i) to reduce duplication when comparing text entities with investigator labels, and (ii) to test whether a simple clustering approach can provide practical utility on small investigative corpora.

For each extracted entity, an embedding vector is generated using the spaCy model. These vectors are grouped through a shallow clustering step. When an entity does not yield a meaningful embedding (for example due to out-of-vocabulary tokens), a fuzzy string matching fallback assigns it to the closest cluster if the token set similarity score is above 85%, picked pragmatically. Otherwise, a new cluster is created.

The result is a set of *canonical entities*, each representing a cluster of surface forms such as *"Per Olsen"*, *"Olsen"*, and *"P. Olsen"*. These canonical entities are used only to stabilise the matching step described below. The ER module is deliberately simple and is not intended as a full entity linking system, but it reduces fragmentation in the text-entity space and makes the comparison with investigator labels more interpretable.

We performed a small exploratory test of three clustering approaches: *agglomerative clustering*, *DBSCAN*, and *k-means*, evaluated with both transformer-based and spaCy-based embeddings. Although the experiment was limited in scope, the results were consistent. Agglomerative clustering produced the most stable and coherent clusters, DBSCAN ranked second, and k-means performed weakest on this type of investigative text. Based on these findings, agglomerative clustering was chosen as the default ER strategy in the pipeline.

3.3.4. Matching

After resolving entities on both the document side and the label side, we compared canonical label entities with canonical text entities to determine which investigator annotations were recovered by each NER model, and vice versa. For every model, a similarity matrix was constructed between the two canonical sets, as illustrated in Fig. 4. Each pair received a combined similarity score s based on the following signals:

s_{emb}: Embedding similarity based on cosine distance between centroid embeddings.

[2] `MarkItDown` by [34] is a Python package that simplifies text extraction from heterogeneous documents.

[3] This study does not aim to guarantee complete recall of all possible entities at the extraction stage. Ensuring absolute coverage is a distinct problem and lies outside the scope of the present work, which focuses on how extracted entities are transformed into investigative relevance.

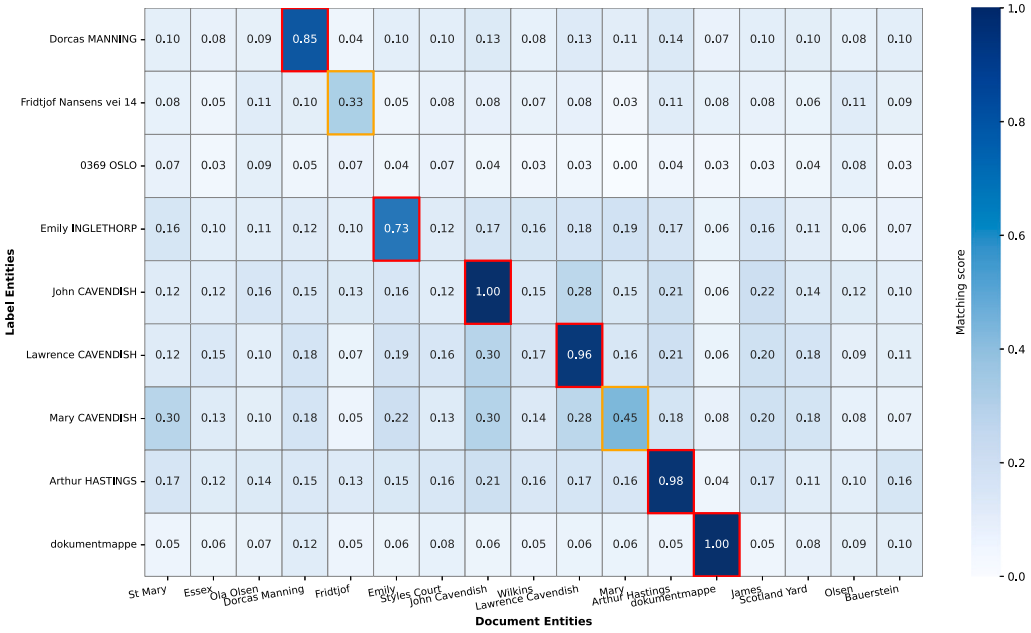


Fig. 4. Similarity matrix between label canonical entities (rows) and document canonical entities (columns) for the Police 3 model. The algorithm performs a *best per label* selection: for each label, the document entity with the highest score above the threshold is chosen as the match, shown with bordered cells. The yellow cells illustrate matches that appear when using the 0.25 threshold but disappear when the cutoff is raised to 0.50, highlighting the effect of stricter matching. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- $s_{\mathbf{ex}}$: Exact canonical-name match, equal to 1 if the canonical names are identical and 0 otherwise.
- $s_{\mathbf{fz}}$: Fuzzy canonical-name similarity, computed using token-set similarity.
- $s_{\mathbf{ov}}$: Exact member overlap, defined as the proportion of identical member strings in the two clusters.
- $s_{\mathbf{mfz}}$: Fuzzy member similarity, defined as the mean token-set similarity across corresponding members.

The final similarity score for a label–document pair was computed as a normalised weighted average of the five signals:

$$s = \frac{\sum w_i\, s_i}{\sum w_i} = \frac{w_{\text{emb}}\, s_{\text{emb}} + w_{\text{ex}}\, s_{\text{ex}} + w_{\text{fz}}\, s_{\text{fz}} + w_{\text{ov}}\, s_{\text{ov}} + w_{\text{mfz}}\, s_{\text{mfz}}}{w_{\text{emb}} + w_{\text{ex}} + w_{\text{fz}} + w_{\text{ov}} + w_{\text{mfz}}} \tag{1}$$

The weights reflected the relative reliability of each signal. Exact canonical-name match and exact member overlap received the strongest weights ($w_{\text{ex}} = 0.6$, $w_{\text{ov}} = 0.6$). Embedding similarity was given a moderate weight ($w_{\text{emb}} = 0.5$), and the fuzzy signals were weighted lower ($w_{\text{fz}} = 0.3$, $w_{\text{mfz}} = 0.2$). Pairs with missing embeddings were scored using the remaining signals only.

Based on these signals we implemented six matching variants. The primary two were *best per label* and *best per text entity*. The *best per label* algorithm identified the highest scoring canonical text entity for each canonical label entity and was used to answer the question *"can we find this label in the document text?"*. The *best per text entity* algorithm examined the reverse direction by checking whether a given canonical text entity was matched by at least one label. This addressed the question *"does the investigator consider this entity relevant?"*.

Thresholds and weights were tuned only on the synthetic case and then kept fixed for all experiments on the real investigative corpus. All match decisions, along with the underlying signal values, were stored in the database for auditability.

3.3.5. Validation

To estimate the theoretical upper bound of how many labels can be recovered from the document text, we manually validated a representative subset of labels. Each label was classified into one of three categories: *directly there*, *indirectly there*, or *not there*. This classification reflects whether the corresponding concept appears verbatim in the document, appears only through paraphrasing or conceptual reference, or is entirely absent from the text. The validation was performed manually by the authors, and the categorisation was carried out independently of the NER output to avoid bias.

To support validation at scale, we developed a custom `Streamlit`-based tool that displays documents, labels, and extracted text entities side-by-side. This tool was used to assign validation categories efficiently and consistently across the dataset. The resulting proportions form the theoretical target for label recovery against which all NER models are evaluated.

3.4. Analytics

All analyses were conducted directly on the structured pipeline output stored in the relational database. Model performance, match rates, and summary statistics were obtained through SQL queries over the extracted entities, labels, and match results. No additional modelling steps were applied.

4. Results

This section presents the results in four parts: (i) label characteristics and theoretical recovery targets, (ii) label recovery across NER models, (iii) document-side entity production, and (iv) canonical entity consolidation after entity resolution.

4.1. Label characteristics and theoretical recovery targets

Table 2 summarises the number of documents and labels for each dataset together with the validation outcomes.[4] Across the two operational cases, 28.8% of labels were classified as not present in the

[4] A 10% validation sample (1 363 of 13 630 labels) provides a statistically robust estimate of the proportion of text-recoverable labels. For a binary outcome (recoverable vs. not), the worst-case standard error of a sample proportion is SE $= \sqrt{p(1-p)/n}$, maximised at $p = 0.5$. With $n = 1\ 363$, this

Table 2
Document and label statistics by case. *Validated* refers to the percentage of labels manually verified. ∉ *Doc.* indicates the share of labels not directly found in the document. *Indirect* refers to labels mentioned only indirectly. *Target* represents the maximum matching percentage theoretically achievable, calculated as the remainder after excluding ∉ Doc. and Indirect labels. Total excludes *The Mysterious Affair at Styles*.

Case	Documents	Labels	Validated	∉ Doc.	Indirect	Target
Case A	1 672	6 743	10.0%	28.5%	12.4%	59.0%
Case B	2 176	6 887	10.0%	28.6%	18.5%	53.1%
Styles	5	49	100.0%	6.1%	6.1%	87.8%
Total*	**3848**	**13,630**	**10.0%**	**28.8%**	**15.4%**	**56.0%**

document text, and 15.4% were present only indirectly. Since these cannot be located reliably by any text-based NER system, the remaining 56.0% represents the theoretical upper bound for automatic recovery.

The synthetic case (*The Mysterious Affair at Styles*) shows broadly similar behaviour: only 12.2% of labels were indirectly present or absent, yielding a higher theoretical target of 87.8%.

4.2. Label recovery across NER models

Table 3 reports the proportion of investigator labels matched to at least one text entity under the two similarity thresholds $M_{0.25}$ and $M_{0.5}$. These can be interpreted in relation to the theoretical targets established in Table 2.

Across the operational cases, the domain-specific models produced the strongest recovery rates. Police 1 and Police 2 matched between 81.9% and 93.9% of all labels under $M_{0.25}$, well above the theoretical upper bound of 56.0%. This is not contradictory: the target reflects only the labels that were *clearly* text-recoverable under manual validation, while the matching procedure can legitimately link paraphrased, partial, or contextually supported mentions outside those strict categories. Similarity-based matching can therefore exceed the theoretical estimate.

Raising the threshold from $M_{0.25}$ to $M_{0.5}$ sharply reduces these effects. Recovery rates for all models drop substantially, isolating the strongest and most reliable text–label correspondences. Under the stricter threshold, Police 1 and Police 2 still achieve the highest recovery rates.

The general Scandinavian baseline (ScandiNER) achieved recovery rates between 60.1% and 83.5%, while the spaCy model, trained on non-investigative news text, remained below the police models but above 65% across all datasets.

In the synthetic case, Police 2 again achieved the highest recovery rate (87.8% at $M_{0.25}$), and the model ranking broadly reflected the pattern seen in the operational cases. This consistency indicates that the domain-specific models capture naming conventions and entity patterns characteristic of Norwegian investigative documents.

4.3. Document-side entity production

The right side of Table 3 reports the number of text entities extracted and the proportion of these that matched at least one label under the two thresholds. The general models (spaCy and ScandiNER) produced far larger numbers of text-level mentions, reflecting their broader type inventories and more permissive extraction strategies. Their lower match proportions indicate that many of these extractions fall outside the scope of investigator-relevant categories.

Police 1 and Police 2 produced smaller, more focused sets of entities with match proportions around 30% at $M_{0.25}$. Police 3, produced more entities than the other police models and achieved higher match proportions than the general models, reflecting its wide ontology-aligned behaviour.

4.4. Canonical entity consolidation

The final column of Table 3 shows the number of canonical entities after entity resolution. Consolidation reduces the number of distinct mentions for all models, particularly for those producing large text-level sets. The relative differences between models remain consistent after canonicalisation: Police 1 and Police 2 produce compact, high-precision canonical sets, while the general models retain substantially larger sets with correspondingly lower proportions of investigator-relevant entities.

5. Discussion and lessons learned

This study evaluated an end-to-end pipeline for automatic entity extraction, resolution, and matching against investigator-generated labels in two operational homicide cases. The results demonstrate that automatic NER systems can recover a substantial proportion of investigator annotations, even when these labels were never designed for machine use and often do not appear verbatim in the underlying documents. At the same time, the analysis reveals systematic differences between what investigators label as important and what NER models extract from the text.

5.1. What the results imply for investigative NER

Three main observations follow from the analysis.

5.1.1. Model success depends on alignment with investigative labels, not domain specificity

Although the police-trained models outperform the general-purpose models, the decisive factor behind their superior performance is not domain specificity in itself. The results show that the critical determinant is whether the model searches for the same types of entities that investigators consider meaningful.

Police 1 and Police 2 make use of an entity taxonomy that closely mirrors investigative annotation practices. These models focus on persons, identifiers, organisations, locations, phones, email addresses, usernames, financial entities, and items. These categories reflect the core investigative questions of identity, communication, movement, and ownership. As a result, when these models identify an entity, that entity is likely to belong to a category that investigators also label.

Police 3 is formally more domain-specific in the sense that it covers a broader investigative ontology, including event categories, weapons, product types, and relational constructs. These categories are rarely annotated by investigators in the present dataset. As a result, Police 3 produces many correct extractions that fall outside the investigator-label space. The resulting gap in recovery rates therefore reflects a taxonomy mismatch rather than model error: the model finds entities it is trained to recognise, while investigators annotate only a narrow subset that is operationally relevant.

The general-purpose models (ScandiNER and spaCy) show the same pattern for a different reason. Their taxonomies originate in newswire and web text, and include categories such as MISC, geopolitical entities, and general products. These are only partially relevant for investigative reasoning. As a result, they extract many entities that have no investigative significance.

This behaviour highlights a conceptual point. Success in investigative NER is not primarily about domain adaptation in the conventional sense. It depends on whether the model's output space reflects the cognitive and operational categories that investigators use. When the

yields SE ≈ 0.0135, giving a 95% confidence interval of ±2.65 percentage points. Applying the finite-population correction for N = 13,630 further reduces the margin to approximately ±2.5 percentage points. This level of precision is more than sufficient for estimating an upper bound such as 56%.

Table 3
Results by case and NER model. *Label match* reports the proportion of investigator labels matched to at least one text entity at the two similarity thresholds $M_{0.25}$ and $M_{0.5}$, comparable to the *Target* values in Table 2. *Document text entities* gives the number of entities extracted by each model and the proportion resolved to at least one label at the same thresholds. *Canonical* shows the number of consolidated entities after entity resolution. Total excludes *The Mysterious Affair at Styles*.

Case	Model	Labels match		Document Text Entities			Canonical
		$M_{0.25}$	$M_{0.5}$	n	$M_{0.25}$	$M_{0.5}$	n
Case A Labels = 6743 Target = 59.0%	ScandiNER	60.1%	35.0%	44,124	45.5%	28.1%	14,878
	Police 1	93.8%	91.0%	9921	30.9%	13.2%	6544
	Police 2	93.9%	92.5%	5207	28.2%	9.1%	3801
	Police 3	63.7%	38.5%	43,718	47.8%	28.8%	14,243
	spaCy	73.6%	57.0%	86,182	30.1%	22.6%	29,847
Case B Labels = 6887 Target = 53.1%	ScandiNER	83.5%	41.4%	70,802	34.4%	15.7%	29,459
	Police 1	84.6%	78.6%	13,254	30.7%	11.7%	9308
	Police 2	81.9%	81.2%	6866	32.3%	10.4%	5323
	Police 3	84.0%	40.8%	65,220	39.2%	19.4%	25,838
	spaCy	84.3%	46.2%	108,140	28.0%	14.3%	52,677
Mysterious Affair at Styles Labels = 49 Target = 87.8%	ScandiNER	73.5%	55.1%	99	56.6%	48.2%	64
	Police 1	85.7%	73.5%	40	60.0%	52.9%	29
	Police 2	87.8%	81.6%	30	66.7%	54.5%	22
	Police 3	73.5%	67.3%	101	51.5%	47.8%	57
	spaCy	65.3%	53.1%	105	48.6%	41.9%	75
Total* Labels = 13630 Target = 56.0%	ScandiNER	71.9%	38.2%	114,926	38.7%	20.4%	44,337
	Police 1	89.2%	84.7%	23,175	30.8%	12.4%	15,852
	Police 2	87.8%	86.8%	12,073	30.5%	9.8%	9124
	Police 3	73.9%	39.5%	108,938	42.7%	23.1%	40,081
	spaCy	79.0%	51.5%	194,322	28.9%	18.1%	82,524

label set is aligned with investigative practice, the model is rewarded for finding the entities that matter, and its output is directly comparable to human annotations. When the label set is misaligned, the model can be technically correct but practically irrelevant.

This distinction is essential for future model development. Instead of asking whether a model is sufficiently domain-specific, it may be more productive to ask whether its taxonomy is aligned with the investigative task. A smaller, focused taxonomy that captures the entities investigators consistently reference may provide more operational value than a larger, more expressive taxonomy trained on a broad investigative corpus.

5.1.2. Investigators select a narrow but meaningful subset of entities

The results confirm that investigators do not attempt to annotate every entity mentioned in a document. Their labels reflect interpretive choices informed by investigative relevance, case hypotheses, and evidential value. Human annotators must also be selective to avoid cognitive overload: exhaustively marking every entity in thousands of documents would be both impractical and counterproductive. As a consequence, large portions of the entities extracted by any automatic NER system, including the police-trained models, are never annotated by humans. This should not be interpreted as model error. It reflects the fact that investigative annotation and automatic NER address fundamentally different tasks.

Investigators selectively capture only those entity mentions that support narrative reconstruction, link analysis, or identity clarification. Typical examples include identifying a suspect, confirming a phone number, or linking a bank account to a person or event. Other textual elements, such as repeated mentions of the same location, generic objects, background characters, or routine administrative details, are usually left unlabelled. These omissions do not represent errors in annotation. They reflect the fact that investigators annotate entities only when doing so contributes to the investigative task.

This has two implications. First, a purely recall-based evaluation can be misleading when applied to investigative data. NER models will always extract entities that investigators did not label, and this difference is structural rather than accidental. Second, investigative workflows benefit from a separation between extraction and relevance assessment. Automatic NER can provide broad coverage, while a dedicated relevance classifier can determine which extracted entities should be surfaced for human review. This pipeline structure mirrors the way investigators think and helps prevent oversaturation with irrelevant information.

5.1.3. Most disagreement reflects structural limits rather than model error

The validation results show that a non-trivial proportion of investigator labels are not directly present in the text. Some labels refer to inferred conclusions, shorthand identifiers, or standardised terminology that does not appear verbatim in the underlying documents. Others are only indirectly supported by the text through paraphrase or conceptual overlap. These distinctions place a natural limit on the achievable matching accuracy of any text-only system.

Because indirect and not-present labels cannot be rediscovered through surface NER, the theoretical upper bound for perfect agreement is significantly below 100 percent. This structural ceiling is visible across all models, including the police models that performed best. The disagreement therefore reflects the nature of investigative annotation rather than limitations of the modelling approach.

This observation has methodological consequences. Evaluation metrics must account for the fact that the annotation space is not text-complete. Measures that assume full text-span alignment, such as strict span-level recall, are inappropriate for investigative corpora. Instead, evaluation should be framed around the recoverable subset of labels and around the utility of NER outputs in downstream tasks such as link analysis or timeline construction. The fact that the police-trained models approach the recoverable limit suggests that the remaining gap is driven primarily by textual absence rather than by NER performance.

Some models appear to exceed the estimated recovery target. This does not imply that they surpass a true upper bound. The target is based on manual validation of a 10% sample of the labels, which identifies only those labels that are clearly recoverable from the text. It is therefore an estimate rather than a strict ceiling. Similarity-based matching can also connect labels that are paraphrased or only partly aligned with the document text, which produces legitimate matches beyond the validated subset. In addition, the matching procedure itself has parameters that influence how strictly or loosely similarity is assessed. This can introduce a small degree of optimistic alignment. Taken together, exceeding the target reflects the behaviour of the matching system and the sampling-based estimate of recoverability rather than any violation of the underlying validation.

5.2. Separation of responsibilities: "find everything" versus "is it relevant?"

A central methodological insight from this study is that automatic NER should not attempt to imitate investigator behaviour. Investigators are not performing linguistic extraction. They are carrying out a relevance judgement driven by hypotheses, evidential needs, and investigative priorities. These tasks differ in training data availability, error tolerance, and operational purpose. Treating them as the same task obscures both. We therefore propose a two-stage architecture:

Find everything A broad-coverage NER model identifies all spans that appear to reference persons, organisations, locations, accounts, identifiers, or other surface-level entities. This component benefits from large, non-sensitive corpora and can be trained or fine-tuned without any access to operational case material. Its goal is coverage, not selectivity.

Is this relevant? A separate model determines which of the extracted entities investigators actually act upon. This component can be trained only on restricted operational data and aligned with investigative priorities. It performs a classification task grounded in relevance, salience, and evidential context rather than in surface form.

This division provides several benefits:

Better generalisation The extraction model remains a stable linguistic tool rather than absorbing investigator habits that differ across units, case types, or regions.

Reduced sensitivity Only the relevance classifier requires access to sensitive case data, simplifying compliance and auditing.

Operational transparency Investigators can influence and interpret relevance scoring without modifying the underlying NER system.

Higher precision The relevance layer can incorporate contextual information, cross-document evidence, case stage, and hypothesis-driven cues that generic NER cannot capture.

Dynamic re-evaluation Because the system retains all extracted entities, relevance can be re-classified as an investigation evolves. An entity initially deemed unimportant may become significant when new information appears, or when investigators shift their working hypotheses.

Our empirical results support this separation. Police 1 and Police 2 extracted fewer entities, but a larger share aligned with investigator annotations. In contrast, spaCy and ScandiNER extracted far more entities, yet a substantial proportion were irrelevant in an investigative sense. Police 3 illustrates a related point. Its taxonomy is richer and closer to the full investigative ontology, and it therefore finds many entities that are potentially useful in principle, even if they are not routinely labelled in the current cases. This is desirable for an extraction layer, which must remain exhaustive with respect to the ontology it is designed to cover. The fact that many of these entities lack corresponding labels does not weaken the argument for a two-stage design. Instead, it highlights that a broad extractor and a relevance classifier serve different but complementary roles.

This indicates that the optimal architecture is not to tune NER to be more selective, but to allow NER to be deliberately broad and to delegate relevance assessment to a downstream model.

5.3. Relevance and entity resolution at multiple levels

A central implication of our findings is that both relevance assessment and entity resolution must operate on more than one level. Document-level processing and case-level processing solve fundamentally different problems, and neither can substitute for the other. Instead, they form a complementary pair.

Document-level relevance and resolution. At the document level, both relevance classification and entity resolution can be applied in a relatively direct and deterministic manner. Entities appear with local context, surface forms are stable, and many relevance decisions can be made based on the immediate linguistic environment. This level is inexpensive to compute, easy to audit, and supports real-time or near real-time workflows. Document-level ER also removes trivial duplication (for example, variant spellings or abbreviations within a single interview transcript), which stabilises downstream analysis.

Case-level relevance and resolution. Case-level reasoning introduces a different set of requirements. Here, relevance is no longer a local linguistic judgement but an analytical inference that depends on accumulation of evidence, cross-document comparison, hypotheses under development, and investigator knowledge. Entity resolution also becomes more complex. Names may change across documents, entities may be referenced obliquely, and cross-document linking requires temporal, relational, and contextual signals that do not exist in any single document.

For this reason, case-level relevance and case-level ER must be allowed to integrate external information such as financial traces, phone metadata, cross-references, or investigator annotations. These steps are necessarily probabilistic, iterative, and subject to revision as the investigation evolves.

Why both levels are required. Relying exclusively on document-level processing would fail to capture cross-document identity and long-range relevance cues. Conversely, relying only on case-level reasoning would be computationally expensive and would lack the stability provided by deterministic document-level extractions. A robust investigative system therefore requires both: a broad, exhaustive, and inexpensive document-level layer, and a more selective, analytical case-level layer that builds on accumulated evidence.

Our results reinforce this distinction. Document-level NER captures a wide field of potential entities, while case-level matching identifies which of these entities actually matter. Similarly, document-level ER removes trivial variation, while case-level ER is necessary to link references across interviews and evidence types. Both layers must therefore be present in any system intended to support real investigative reasoning.

5.4. Future work

Future work will extend this study in three main directions. First, the next stage is to develop a supervised relevance classifier using the dataset produced in this project. The results clearly show that linguistic extraction and investigative selectivity are distinct tasks. A dedicated relevance model will allow this separation to be implemented in practice. Such a model should rely primarily on document-level signals, since document-level relevance is fast, cheap, and largely deterministic. Case-level information can still enrich relevance scoring further downstream in the analysis pipeline, once cross-document context and evolving investigative priorities become available.

Second, the entity resolution component should be expanded beyond the initial clustering experiment. With larger datasets, it becomes possible to evaluate span embeddings more systematically, test hybrid fuzzy–vector approaches, and design multi-stage candidate-ranking procedures. This extended resolution step also forms the basis for relationship extraction. A practical system will need to capture simple

co-occurrence patterns, rule-based syntactic relations, and more subtle links that require transformer-based interpretation. These relations are essential inputs to downstream investigative reasoning.

Third, the long-term trajectory is to construct a functional knowledge graph over complete investigations. The extraction module and the relevance classifier together provide the inputs for graph construction, entity linking, pruning, and hypothesis-testing mechanisms. A graph layer enables cross-document reasoning, retrieval-augmented question answering, historical reconstruction of events, and the identification of entities or statements that warrant further attention.

Taken together, these efforts move the work from evaluating NER behaviour towards building a full investigative analysis pipeline. A relevance classifier, a strengthened resolution module, and a graph-based reasoning layer form the natural and technically grounded next steps.

5.5. Limitations

This study is based on two operational homicide cases and one synthetic case, which limits generalisability to other investigative domains. The labels used as ground truth were not designed for machine learning and often do not correspond directly to textual spans. Many are indirect or conceptual, which constrains the theoretical upper bound for any text-only system.

The documents also contain a large amount of boilerplate and structural content. General-purpose NER models tend to extract entities from this material, while investigators rarely annotate it. This affects document-side precision and should be considered when interpreting extraction volume.

Similarity weights and thresholds were tuned only on the synthetic case and then held constant. While methodologically consistent, this may not be optimal for cases with different writing styles or entity distributions.

Finally, the entity resolution module is intentionally simple. It stabilises surface variants but does not perform full semantic disambiguation or cross-document linking. A more advanced ER module is needed to support richer case-level analysis.

6. Conclusion

This study examined how well automatic NER can rediscover the entities that investigators annotate in real criminal cases. By applying five models to 3848 operational documents and comparing their outputs to 13,630 investigator labels, we showed that a substantial share of labels cannot be recovered from text alone. These structural limits define a realistic ceiling for any text-only system.

Across models, the strongest performance came from those whose entity taxonomy aligns with what investigators actually label. At the same time, the broader extractions produced by general-purpose models underline the need for a downstream relevance component. Taken together, the results support a two-stage architecture: a broad, linguistically driven NER model followed by a relevance classifier that reflects investigative priorities.

The study also highlights the value of combining document-level and case-level reasoning. Document-level processing provides fast and deterministic signals, while case-level aggregation allows relevance to evolve as the investigation progresses.

The work establishes a baseline and a methodological template for future research. Next steps include training a relevance classifier on the prepared dataset, extending entity resolution, and incorporating relation extraction and knowledge-graph construction to support richer case analysis.

CRediT authorship contribution statement

Johannes Pippidis Lorentzen: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Mads Skipanes:** Validation, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This project is financed by the Norwegian Research Council (project numbers 333898), Kripos, DNB and NTNU.

References

[1] R. Al-Zaidy, B.C. Fung, A.M. Youssef, F. Fortin, Mining criminal networks from unstructured text documents, Digit. Investig. 8 (3–4) (2012) 147–160.
[2] A. Louis, A.P. Engelbrecht, Unsupervised discovery of relations for analysis of textual data, Digit. Investig. 7 (3–4) (2011) 154–171.
[3] M. Skipanes, G. Demartini, K. Franke, A.B. Nissen, Information analysis in criminal investigations: methods, challenges, and computational opportunities processing unstructured text, Polic.: A J. Policy Pr. 19 (2025) paaf005.
[4] J. Partridge, F. Zaghloul, Policing the data: Can data analytics help law enforcement? J. Inf. Technol. Teach. Cases 15 (1) (2025) 90–94.
[5] M. Skipanes, T.E. Jørgensen, K. Franke, Advancing knowledge discoveries in criminal investigations with semantic textual similarity, Frontiers Artificial Intelligence Appl. 379 (2023) 269–274, http://dx.doi.org/10.3233/FAIA230973.
[6] D. Jurafsky, J.H. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2025.
[7] E.F. Sang, F. De Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, 2003, ArXiv Preprint Cs/0306050.
[8] Z. Hu, W. Hou, X. Liu, Deep learning for named entity recognition: a survey, Neural Comput. Appl. 36 (16) (2024) 8995–9022.
[9] M. Schraagen, M. Brinkhuis, F. Bex, et al., Evaluation of named entity recognition in dutch online criminal complaints, Comput. Linguist. Neth. J. 7 (2017) 3–16.
[10] M. Skipanes, N. Pratama, K. Porter, G. Demartini, Fast synthetic data generation for case-specific entity extraction in criminal investigations, in: Proceedings of the Digital Forensics Doctoral Symposium, DFDS '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–8, http://dx.doi.org/10.1145/3712716.3712719.
[11] O. Binette, R.C. Steorts, (Almost) all of entity resolution, Sci. Adv. 8 (12) (2022) eabi8021.
[12] O. Bender, F.J. Och, H. Ney, A maximum entropy approach to named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning At HLT-NAACL 2003, Association for Computational Linguistics, 2003, pp. 160–163.
[13] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann, 2001, pp. 282–289.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.
[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[16] K. Porter, B. Skattør, Mapping NLP techniques to investigations and investigative interviews, in: Z. Geradts, K. Franke (Eds.), Artificial Intelligence (AI) in Forensic Sciences, first ed., Wiley, 2023, http://dx.doi.org/10.1093/police/paaf005.
[17] M. Chau, J.J. Xu, H. Chen, Extracting meaningful entities from police narrative reports, in: Proceedings of the 2002 Annual National Conference on Digital Government Research, 2002, pp. 1–5.
[18] J.R. Johnson, A. Miller, L. Khan, Law enforcement ontology for identification of related information of interest across free text dcouments, in: 2011 European Intelligence and Security Informatics Conference, IEEE, 2011, pp. 19–27.
[19] E. Casey, S. Barnum, R. Griffith, J. Snyder, H. van Beek, A. Nelson, The evolution of expressing and exchanging cyber-investigation information in a standardized form, in: Handling and Exchanging Electronic Evidence Across Europe, Springer, 2018, pp. 43–58.

[20] A.M. Ørke, Giving the Police a Head Start: Norwegian Named Entity Recognition Dataset and Model Development for Investigative Purposes (Master's thesis), Norwegian University of Science and Technology (NTNU), 2023, URL https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3103088.
[21] A. Nissen, Informasjonsbehandling i store straffesaker, 2018.
[22] L. Gianola, Traitement automatique des langues et linguistique de corpus pour la reconnaissance d'entités en analyse criminelle, Rev. Int. de Criminol. Police Tech. Sci. 3 (2021) 363–382.
[23] M. Bérubé, F. Fortin, O. Péloquin, La reconnaissance d'entités nommées: une démarche. prometteuse pour la détection automatisée de liens dans les documents d'enquête policière, Criminol. Forensique et SécuritÉ 1 (1) (2023) 3349.
[24] C. Batini, V. Bellandi, P. Ceravolo, F. Moiraghi, M. Palmonari, S. Siccardi, Semantic data integration for investigations: Lessons learned and open challenges, in: 2021 IEEE International Conference on Smart Data Services, SMDS, IEEE, 2021, pp. 173–183.
[25] F.B. Rodrigues, W.F. Giozza, R. de Oliveira Albuquerque, L.J.G. Villalba, Natural language processing applied to forensics information extraction with transformers and graph visualization, IEEE Trans. Comput. Soc. Syst. (2022).
[26] H. Henseler, H. van Beek, ChatGPT as a copilot for investigating digital evidence, in: LegalAIIA@ ICAIL, 2023, pp. 58–69.
[27] M. Skipanes, J.P. Lorentzen, S.Y. Yayilgan, Understanding the role of case management systems in criminal investigations, 2025, Manuscript under second-round review.
[28] K. Finlay, M. Mueller-Smith, J. Papp, The criminal justice administrative records system: A next-generation research data platform, Sci. Data 9 (1) (2022) 562.
[29] X. Xing, P. Chen, Entity extraction of key elements in 110 police reports based on large language models, Appl. Sci. 14 (17) (2024) 7819.
[30] M.H. Amirhosseini, H. Kazemian, M. Phillips, A graph-based method for identity resolution to assist police force investigative process, J. Cyber Secur. Technol. 9 (2) (2025) 127–150.
[31] D. Robinson, The use of reference graphs in the entity resolution of criminal networks, in: Pacific-Asia Workshop on Intelligence and Security Informatics, Springer, 2016, pp. 3–18.
[32] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, H. Liu, Large language models for data annotation and synthesis: A survey, 2024, arXiv preprint arXiv:2402.13446.
[33] A. Christie, The mysterious affair at styles, 1920.
[34] Microsoft, MarkItDown, 2025, URL https://github.com/microsoft/markitdown, GitHub repository; (Accessed 21 October 2025).
[35] D.S. Smart, ScandiNER: `saattrupdan/nbailab-base-ner-scandi`, 2024, URL https://huggingface.co/saattrupdan/nbailab-base-ner-scandi.
[36] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength natural language processing in python, 2020, http://dx.doi.org/10.5281/zenodo.1212303, URL https://doi.org/10.5281/zenodo.1212303.
[37] Explosion, spaCy Norwegian Bokmål pipeline: `nb_core_news_lg`, 2025, URL https://spacy.io/models/nb#nb_core_news_lg.