MULTI-CASE: A Transformer-based Ethics-aware Multimodal Investigative Intelligence Framework

Maximilian T. Fischer, Yannick Metz, Lucas Joos, Matthias Miller and Daniel A. Keim

***Abstract*—AI-driven models are increasingly deployed in operational analytics solutions, for instance, in investigative journalism or the intelligence community. Current approaches face two primary challenges: ethical and privacy concerns, as well as difficulties in efficiently combining heterogeneous data sources for multimodal analytics. To tackle the challenge of multimodal analytics, we present MULTI-CASE, a holistic visual analytics framework tailored towards ethics-aware and multimodal intelligence exploration, designed in collaboration with domain experts. It leverages an equal joint agency between human and AI to explore and assess heterogeneous information spaces, checking and balancing automation through Visual Analytics. MULTI-CASE operates on a fully-integrated data model and features type-specific analysis with multiple linked components, including a combined search, annotated text view, and graph-based analysis. Parts of the underlying entity detection are based on a RoBERTa-based language model, which we tailored towards user requirements through fine-tuning and published as open-source. An overarching knowledge exploration graph combines all information streams, provides in-situ explanations, transparent source attribution, and facilitates effective exploration. To assess our approach, we conducted a comprehensive set of evaluations: We benchmarked the underlying language model on relevant Named Entity Recognition (NER) tasks, achieving state-of-the-art performance. The demonstrator was assessed according to intelligence capability assessments, while the methodology was evaluated according to ethics design guidelines. As a case study, we present our framework in an investigative journalism setting, supporting war crime investigations. Finally, we conduct a formative user evaluation with domain experts in law enforcement. Our evaluations confirm that our framework facilitates human agency and steering in security-sensitive, AI-supported analysis processes while addressing ethical and privacy concerns and providing much-needed analytical capabilities.**

***Index Terms*—Intelligence analysis, communication analysis, investigative journalism, case study, ethical, evaluation, multivariate, multimodal analytics, multimedia analysis, visual analytics.**

I. Introduction

AI-DRIVEN models have gained wide popularity over the last few years and have been applied successfully in numerous fields, such as natural language processing (NLP), computer vision, or predictive analytics. Given this general trend, AI models are increasingly needed [1] and deployed in operational intelligence solutions [2], [3]. Corresponding application domains, such as investigative journalism [4], [5] or the intelligence domain [2], [6], [7], [8], are particularly interesting due to their unique set of distinct challenges. Intelligence analysts often face the task of combining numerous, heterogeneous pieces of intelligence, often tainted with uncertainty and conflicting information, forming an incomplete picture. As discussed in previous work [9], the first set of challenges in this regard is related to **ethical** [10] and **privacy concerns** [1] due to the sensitive nature of the data and operations involved [11] and the high stakes in case of errors [12], [13]. Simultaneously, these domains offer opportunities for increasingly automated, tailored systems to deal with incomplete and tainted information. This is particularly the case for **heterogeneous** and **multimodal analytics**, a second area in which existing systems often lack in functionality [14], [15].

The analysis of **individual modalities** in isolation—like network structure of the participants, named entity detection on the content, or time series analysis of the individual message intervals—often comes with limited views on the underlying information with consequences for the derived intelligence. Not considering these aspects can reduce trust in AI systems, favor prejudices and mistakes, and also lead to legal consequences. Further, isolated analysis requires human knowledge and intervention to semi-manually find hidden cross-matches between the modalities—a task where computational support can be highly effective, reduce domain discontinuities, and place less additional workload on the users [14]. This becomes even more important when users are no machine learning experts, thus sometimes having unrealistic expectations or misplaced trust in the systems [1], [9]. This can be the case for (business) intelligence analysts or investigative journalists, after which we modeled a case study (see Section V-A).

This study is based on widespread **tasks** in intelligence, identified by the UNODC [16], which aims to answer the typical six questions: *Who? What? How? Where? Why? When?* Based on these six questions, the UNODC authors identify three common analysis tasks and methods that typically enable the answering of these questions in relevant investigations: (1) *link analysis*: searching and identifying relationships between specific entities such as persons or organizations, but also objects, locations, or events, (2) *event analysis*: correlating actions or locations alongside their timeline order, (3) *flow analysis*: understanding the connectedness as well as cause and result, for example, the flow of commodities (geolocation for physical goods or transfers of money) or the propagation of knowledge. Other tasks described in the report involve the identification of activities, frequencies, or general data correlations. These tasks can be primarily achieved through four main methods: (a) keyword and semantic-based searches on text or transcripts to understand the context or find entities, (b) (social) network analysis to find connections and relations, (c) meta-data-filters to restrict, for example, locations, and

M. T. Fischer, Y. Metz, L. Joos, M. Miller, and D. A. Keim are with the University of Konstanz. E-mail: {max.fischer, yannick.metz, lucas.joos, matthias.miller, keim}@uni-konstanz.de.

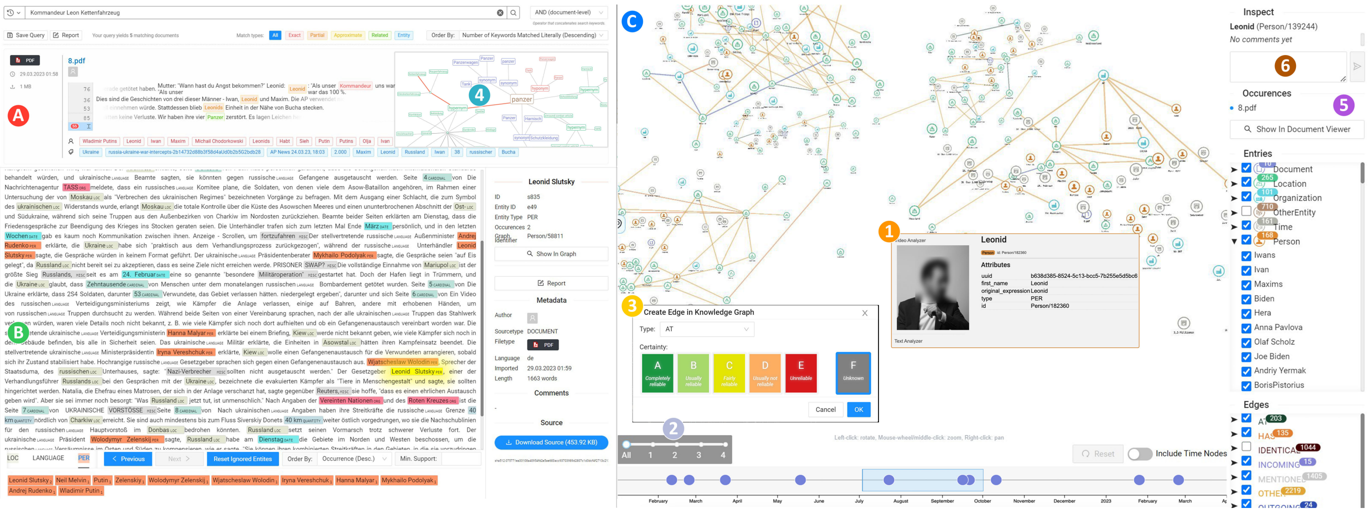

Fig. 1. **MULTI-CASE: A holistic visual analytics framework tailored towards ethics-aware and multimodal intelligence exploration.** Built upon a fully-integrated data model, it features type-specific, graph-based analysis through individual models with multiple, linked components: Ⓐ a combined ontological search and result interface, Ⓑ an interactive textual view, and a Ⓒ knowledge graph interface. Additional components (not shown) include more specialized modules like video or audio analysis. The interface facilitates ① in-situ contextualization across modalities, ② graph neighborhood explorations, ③ relevance scoring for accountability and oversight, ④ transparent source explanations, ⑤ integrated navigation, and ⑥ collaborative user participation.

(d) time-series analysis, for example, to identify particular communication patterns. However, these modalities should not be considered to work in isolation but contribute individual perspectives for corroborating, enhancing, and setting each other in context. For example, to attribute war crimes in our case study (see Section V-A), our journalist Alisa leverages semantic analysis, geolocation, link-analysis, and time-correlation together with several other methods to achieve her objectives.

Our **objective** is to tackle the existing shortcomings in ethical and multimodal analysis for intelligence by presenting a framework for holistic communication analytics. Many specific solutions have been proposed, but the integration and combination have received less attention. In previous work [9], [15], we have detailed the data and problems faced in intelligence analytics: the need for heterogeneous data analytics capability due to the diverse set of intelligence received. The different data types and scenario stakeholder groups like data subjects, software providers, civil society, and governmental authorities with their different branches with all their conflicting interests. Their requirements and tasks, which we also revisit below, as well as the benefits and possible designs of visual analytics applications.

Our contribution is not intended as a fully-fledged analytics system but as an exemplary *framework* for a holistic, multimodal approach to intelligence and its assessment. Therefore, we dedicate significant time towards a comprehensive *evaluation* (see Section V), encompassing multiple perspectives, i.e., ethical aspects, capabilities, and practical considerations through use cases and expert studies.

Based on lessons learned in previous work [9], [14], [15], we aim to enhance the analytical capabilities in semi-automated digital intelligence analysis, making the following **contributions**:

- MULTI-CASE, an *integrated* **visual exploration framework** (see Fig. 1) tailored towards ethics-aware *multimodal* intelligence analytics in investigative journalism or criminal investigations.
- A RoBERTa-based NER **transformer model**, derived by fine-tuning on GottBERT [17] alongside intelligence-specific training data, which we both open-sourced at osf.io/eap4r.
- An extensive **case study** showcasing MULTI-CASE in the context of *war crime investigations* together with a **classification assessment** of its **capabilities** [15] and **ethics design** [9].
- A formative **expert evaluation** with eleven domain experts in different law-enforcement areas, validating the approach's advantages and highlighting areas for further improvement.

With this contribution, we fill a gap in bringing state-of-the-art performance to applications by providing an explainable visual exploration framework for multimodal intelligence analytics. We consider our contribution primarily in the combination of existing visualization and visual analytics methodologies suitable for this domain and their detailed assessment in the context of the unique challenges faced. Thereby, we aim to provide more insights into the often opaque workings in the intelligence domain, furthering research and a critical discussion.

II. Related Work

The research on **multimodal visual intelligence** analysis is sparse. While there is significant literature on intelligence analysis in general [16], [18] and some requirement studies for general intelligence analytics tools exists [15], [19], [20], actual tool descriptions are rare. If a paper evaluates an actual approach, their findings primarily focus on user acceptance while ignoring capabilities or interactive visualizations since tools are often classified and not even named publicly [21].

Research on some of the underlying techniques itself, for example, classical **Named Entity Recognition** (NER) as the foundation for comprehensive tasks like entity linking, is much more common. Techniques evolved over time from using rule-based to more statistical systems [22]. Traditionally, NER relied on annotated corpora, which posed challenges for domain transfer and new label tasks, with brittle results [22], [23]. However, with the advent of deep learning-based approaches, such as BERT [24], the landscape has changed, and transfer learning (i.e., adapting pre-trained models to shorten training times for new tasks) can cope with much smaller amounts of annotated text. This has significantly improved the adaptability and efficiency across various domains and tasks, making knowledge transfer and few-shot labeling easier [24], [25], which can be leveraged in investigative tools.

Similarly, advances in **ethical design** [9], [26], like the concept of providing guidance [27], visualizing hidden uncertainties [28], or ensuring provenance [29] as well as **privacy considerations** [9], like selective masking [30], federated learning [31], or data perturbation [32] have been made. Also, **insular solutions** like Pajek [33] for social network analysis, Maltego [34] or InSight2 [35] for link analysis, or Cosmos [36] for semantic text analysis exist but do not combine modalities.

Within the visualization community, **multimodal multimedia analysis** [37] can be considered partly related: Several approaches have been proposed to consider different aspects of multimedia content simultaneously, like the presentation styles and techniques [38], the emotional coherence [39], or the automation of explicit content through video moderation [40]. While these approaches propose valuable insights into how (primarily visual) media can be analyzed and set into context, many of the approaches target very specific applications, and very few in this domain truly support a holistic approach to analyzing *generic* pieces of intelligence, which also includes text-based information. Further, Zahalka and Worring presented a pathway to comprehensive multimedia analytics, detailing a general four-tiered multimedia analytics model and discussing it alongside how it may support addressing the semantic and pragmatic gap encountered in actual systems [37]. This follows a similar overall direction as our research, however, with one particular difference: The model is applicable in general for the analysis of multimedia data and also with a particular focus on such data, for example multimedia collections of images. While some aspects overlap, these collections of images do not necessarily have a underlying storyline, may come from any collection mechanism (e.g., underwater camera), an the model primarily focuses on a multimodal analysis of multimedia with additional metadata (e.g., annotated text or features). Our approach instead focuses primarily on communication between humans, emphasizing much more the interactive aspects of the information exchange via various modalities over time.

The research on leveraging **visual analytics for intelligence applications** [41], [42], [43], [44] had its prime in the mid-to-late 2000s, with frameworks such as VIM [45] or Jigsaw [46]. Both primarily focus on text documents (and not so much multimedia), and only a few approaches [14] were proposed later on. Therefore, this area seems to be one of those few domains where commercial research has outpaced academic, scientific research for now.

In the context of actual usage—also for commercial systems—we surveyed related communication analysis systems [15], where we identified four publicly known intelligence systems in wider use: DataWalk [47] and Nuix Discover / Investigate [48] are sometimes used, while the market leaders are IBM i2 Analyst's Notebook [49] along with Palantir Gotham / Foundry / Meta-Constellation [50]. While they cater to government applications, parts are commercially available and are used by international banks, advertisers, manufacturers, telecommunication providers, media organizations, and NGOs [50].

To our knowledge, no new visual analytics approaches to intelligence have been publicly proposed since our recent survey of AI-driven intelligence applications [15], also available as an **interactive browser** at https://communication-analysis.dbvis.de. Regarding practical usage, the ongoing shift from IBM i2 to Palantir seems to accelerate. Palantir's solutions (in particular Meta-Constellation) are also employed effectively [51] by Ukraine in its defense against Russia in coordinating their military.

The **academic research** on this topic has been falling short, with problematic consequences for accountability and oversight, which has also been realized by some key stakeholders. For example, in the European Unions Horizon 2020 funding period alone, projects such as ASGARD (700381), MAGNETO (786629), STARLIGHT (101021797), COPKIT (786687), and AIDA (883596) (some still ongoing) have been funded, although preliminary results show insular capabilities. For the upcoming Horizon Europe funding period, several calls have been proposed (e.g., HORIZON-CL3-2023-FCT-01). With slight deviations, they all aim to increase analytical big data capabilities for law enforcement. In the US, similar research is often conducted by national laboratories but mostly remains classified.

While many visualization approaches can be leveraged for intelligence, only few consider the combination of challenges faced in this particular domain, including the inherent uncertainty and inter-modality, while even fewer evaluate them consistently and publish the results, which is the goal of this work.

III. Methodology: Model Development

In previous work [14], we have presented a matrix-based, holistic communication analysis framework through semantic zooming. As our studies have shown, however, despite the potential benefits in scalability, matrices are uncommon for many analysts, which are used to graph- and relationship-based visualizations. Further, semantic zooming is space-limited in the amount of context information in the upper layers. We, therefore, aim to explore an *orthogonal design*, with *two key advancements*: (a) Following a similar modular approach, we leverage a more powerful **fully-integrated data model** (structuring and relating the intelligence information pieces)

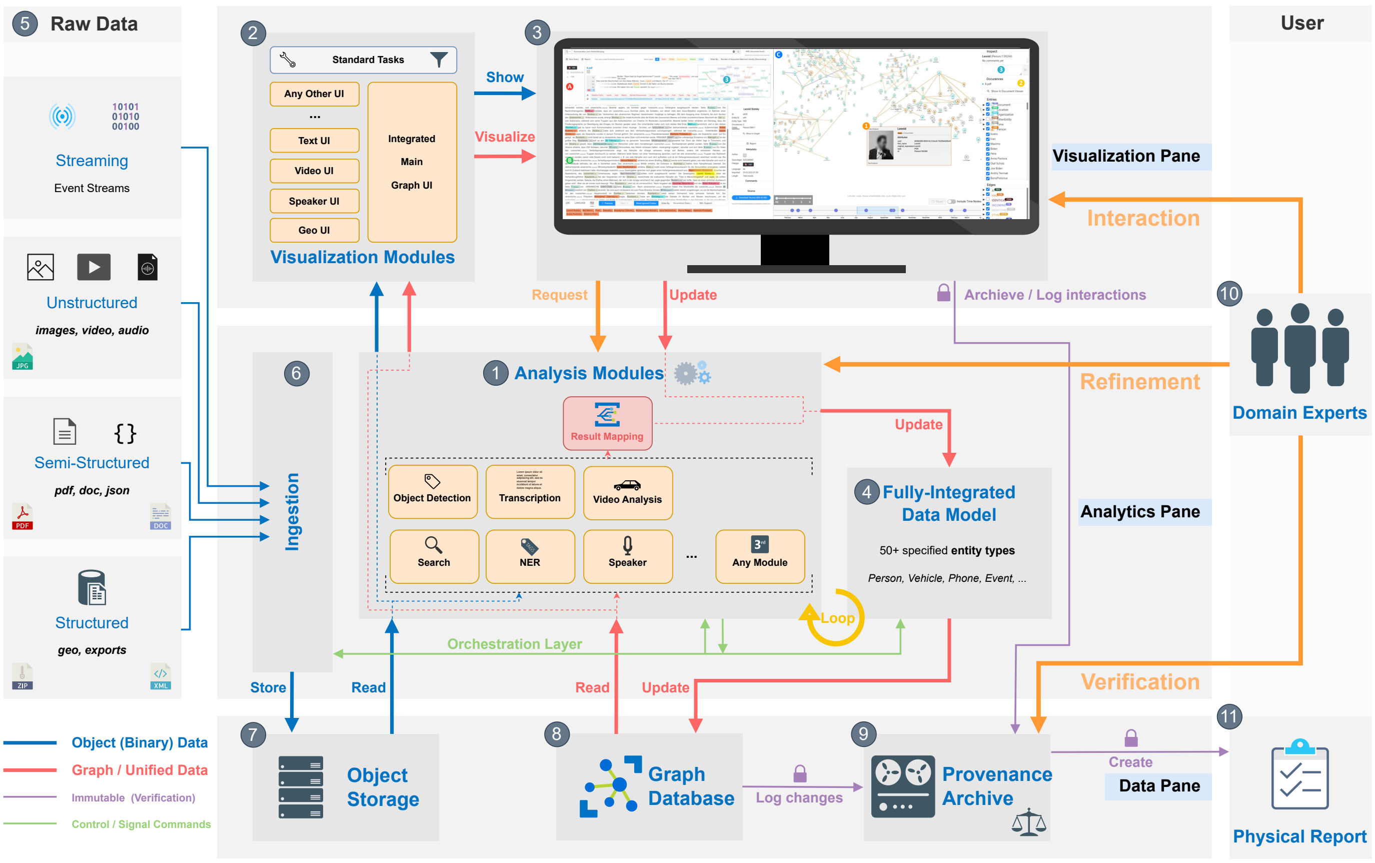

Fig. 2. **High-level architecture of MULTI-CASE**, highlighting the main components ① - ⑨ , like the ① analysis modules or the ③ graph UI with ② visualization modules, the ④ fully-integrated graph data model, as well as different data paths and user handling, like the — analytics workflow. A detailed description is provided in Section IV.

that also supports multimodality. (b) Instead of matrix-based semantic zooming, we use a **graph-based overview** with several linked views and integrated specialized views.

This decision is based on the task descriptions and requirements described in the UNODC report [16] described above, as well as feedback from several domain experts in law enforcement, which state the following three **user requirements** for such a framework: (1) A centralized, multimodal platform for collaborative case working. (2) Assistance in labor-intensive tasks such as big data analytics. (3) Transparency and reliability.

This reflects their need to *work collaboratively* on a case together with their colleagues on larger investigations, needing to share results or to leverage knowledge generated by colleagues investigating specific aspects of a case by collaboratively working on a shared data space and being able to access the information in-situ. Due to the sheer volume and sometimes repetitive tasks, support by automation and AI is considered essential while being reliable and understandable. All the while, the analysis steps taken need to be transparent and reproducible for accountability. Guided by these overall principles, we further justify individual design decisions and capabilities while describing the system design in Section IV.

One central aspect of intelligence analytics is the analysis of communication [14]. However, common international NER labeling schemes (e.g., PER, ORG, LOC, OTH) often do not meet the specialized requirements for investigations since they are too ambiguous and not specialized enough, requiring more narrow tag categories [52], [53]. In practice, specialized **NER model development** for semantic understanding is still challenging, with many pitfalls, although the attention-based transformer architecture [54] has significantly increased the accuracy compared to previous neural models. Therefore, as part of this work, we track the necessary steps for training and deploying transformer models, including interactive tools for labeling, while also highlighting major lessons learned. The necessary steps range from choosing a suitable base model, preparing representative training data, then training and evaluating, to finally supervising and validating the model in deployment and adapting it in the face of changing language patterns, terms, or requirements. As a result, we provide a strong baseline NER transformer model with a large set of relevant entity labels to simplify future applications. For the underlying language model, we considered existing models from the Huggingface transformers [55] library based on evaluation performance on the GermEval14 dataset [56], a well-known dataset for German NER recognition. For German natural language processing, we considered two language models: The RoBERTa-based GottBERT [17] and *BERT-base-german-cased* [57] based on the original BERT transformer architecture [24]. Additionally, we chose a strong multi-lingual baseline (XLM-RoBERTa) [58].

In general, the creation of specific training datasets, for example, through **labeling** of domain-specific datasets, is often tedious and error-prone. Therefore, we implemented an interactive labeling tool that is compatible with the MULTI-CASE framework, allowing us to label and subsequently review a given document collection on a large scale, facilitating the easy creation of ground truth training datasets in specific domains, like intelligence. This is particularly relevant in our application because it utilizes a large set of custom-named entity labels for domain-specific analysis. Many non-English models only provide standard categories like *PERSON*, *LOCATION*, *ORGANIZATION*, and *MISC*. However, based on expert feedback, custom categories like *EVENT* or *PRODUCT* and more fine-grained time and numeric labels were introduced, with the full list shown in Table I. We provide an enhanced, re-tagged version of GermanNER alongside our model at osf.io/eap4r.

For the **training**, we apply a train/validation/test split of 70/15/15 of the full mixed dataset (domain-specific and re-tagged corpus data). We train each baseline model with Adam [59], weight decay [60], and 0.1 dropout. We also experimented with a slanted triangular learning rate (i.e., using a warm-up and linearly decaying learning rate) [61] and found a slight positive effect on final performance. Early stopping was implemented based on the validation F-score with a number of patience steps of 10. Fine-tuning of all models was performed on a single RTX4000 GPU. We report the full set of hyperparameters and additional results at osf.io/eap4r.

After training, we evaluate the model performance alongside other base models on a held-out test set and describe the results in Section V-D. We also note the recent advances by Large Language Models (LLMs), which can drastically improve specialized NER-tagging through zero- or few-shot learning, in the outlook in Section VI-B.

IV. System Design

The proposed architecture for our framework and the fully-integrated graph data model described in the following is shown in Fig. 2. When necessary, we also detail the expert reasoning and the ethical considerations behind individual design decisions while also referring to Sections III and V-E as well as V-B for further discussions on these topics. The guideline numbers for the ethical and privacy reasoning (e.g., C1-6, R1-5, A1-6) refer to the nomenclature established in previous work [9].

Overall, the system consists of individual plugins ❶ Analysis Modules for specific analysis tasks and data types, a ❸ Main Graph-based UI together with specialized ❷ Visualization Modules (e.g., text analysis or video-analysis) for a web-based exploration. This fulfills the demand by experts to be capable of specialized analysis that interfaces with an overall case working framework. Similarly, the heavy computations are run on a centralized server, while the interface nowadays is a standard web-based approach running on a regular (or thin) client. One key aspect of the overall system is the ❹ Fully-Integrated Data Model stored in a ❽ Graph Database, which acts both as a conceptual abstraction layer between modules and a central source of shared knowledge. This enables the experts to work on a consistent data set in an integrated environment and not lose information compared to switching between applications, increasing Efficiency (A4) while addressing the working together of machines and users (C5). Supporting roles fall to the ❼ Object Storage to store any input and intermediate data and the ❾ Provenance Archive as a revision-safe storage, which is considered essential for Opacity (C3) and Accountability (C6). The ❿ domain experts can communicate with the system by interacting with the visualization, forming a collaborative Human-Machine-Configuration (C5), refine the display through analysis parameters, as well as verify the results, which increases understanding and fosters trust and works against Lack of Accountability (R1), while enabling Human Oversight (R5) and also facilitating a critical reflection (R4). This verification is available both in the interface and in a ⓫ physical report, which the experts still need to document their findings in a structured way.

A. Data Model

Diverse types of ❺ Raw Data are supported, ranging from unstructured data (images, video, audio), over semi-structured documents (e.g., PDF documents), to structured data types (like geolocation tracks or exports), as well as streaming data. The needs of the domain experts naturally vary here depending on their organization and tasks, but typically the first two types are the most common ones. The input is only limited by the plugin analysis modules. When data is ❻ ingested, it is stored in the ❼ Object Storage. Based on the input type, the Orchestration Layer selects one or more analysis modules for knowledge extraction, for example, NER for text documents. The main results are mapped to the ❹ Fully-Integrated Data Model stored in the ❽ Graph Database. For example, for NER, this could be the detected *entities*, like persons, location, or dates, as well as their *relations*, while for video analysis, an object like a car along its properties and a relationship to time and location. Two aspects are of primary importance:

(1) the **data model** ideally has to be as mutually exclusive and collectively exhaustive as possible. The data model was designed with several domain experts and generalized from existing case models like IMP (Information Model Police). In our case, we arrived at 50+ hierarchical *entities* (graph nodes) and 10+ *relationship* types (graph edges), trying to find the right balance between a generic data model and enough specialization. While a very generic data model allows for the reflection of virtually all analysis results, the automatic conclusions, connections, and information enrichment in such a case can remain very limited. In contrast, a highly specialized data model allows to reflect on the findings with high precision and enables many automated conclusions. However, it always poses the danger of being too specified (i.e., available properties on a type) to capture all relevant information. Indeed, the principle design is flexible, subject to change, and can be adapted by adding more specialized entities or data fields. Analysis modules are change-agnostic if the entities and attributes they work with are untouched. In our case, we derived everything from a root *Thing*, with *Entity*, *Event*, *Datetime*, *Location*, and *Document* as the first hierarchical layer, each having further subtypes (e.g., *Person* or *PhoneCall*). For example, a *Timespan*, as a subtype of *Datetime*, represents a specific time range and can be related to a *PhoneCall* via a relationship, which in turn may be related to specific phone numbers, which again might be related as belonging to actual persons. Attributes for each entity store associated information. Through *relationships*, one can also model source attribution (source document and analysis module) and ③ confidence scores, e.g., based on the 6x6 intelligence scoring [16], which many analysts are well familiar with, strengthening Literacy (A5). This scoring can have an influence on automatic decision-making: when certainties are considered by algorithms, this can support working towards Preventing Automated Inequality (R3) and limit Exaggerated Expectations (C4) and Discriminatory bias (C1) through manual priming. Simultaneously, the opposite could also be true, where the system warns a user of inherent prejudice evident in analysis choices.

(2) The data model allows a structured information **exchange** and also **information enrichment process** between modules, which the experts consider essential. Updates of the data model can trigger subsequent runs of other analysis modules when they have signed up for specific creations/updates: for example, an imported audio file might be analyzed first by a speaker detection (with the creation of a specific audio entity), then by a speech-to-text transcription (with a text entity), and then by a NER process, which can result in an enrichment of the graph with the conversation content through multiple entities (e.g., persons, location, or times). All changes (creations, updates, hidings) in the graph data are logged via a write once, read many ⑨ Provenance Archive.

B. Component Integration

The individual ① Analytics Modules like NER or transcription are designed as plugins and can be flexibly combined depending on the analytical needs, allowing for Customization (A6) and ensuring User Agency (A1) of the experts. In this work, we primarily focus on the search and NER modules as an exemplary prototype developed by us, while other modules are provided as open source (e.g., transcription via Whisper [62]) or by commercial partners. During startup, the modules register themselves, their supported data types for ingestion, and the graph change listeners via the Orchestration Layer. Further, each analytics module can register custom *context actions* (e.g., show similar persons) and *preview handlers* (e.g., picture or video player), which are integrated into the ③ Main Graph UI, allowing for a tight coupling between the UI and individual modules functions in ② specialized UIs, supporting the mental mapping of the experts.

C. Interfaces and Interaction Principles

The interfaces are web-based, and the provided views are **tightly coupled** and **inter-linked**, strengthening the Human-Machine-Configuration (C5) and the User Agency (A1) through Opacity (C3). Entities are consistently mapped via the unified, fully-integrated data model, allowing for the enrichment of information within the main graph-based overview and across views.

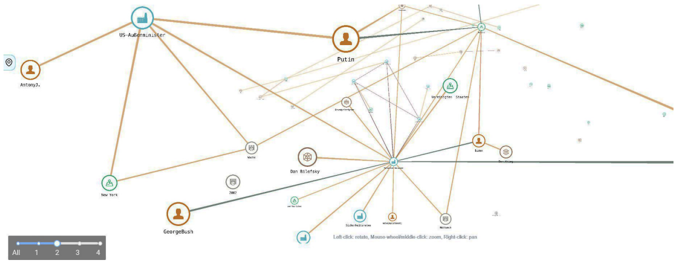

Fig. 3. **Neighborhood exploration** ② , acting as a magnifying spotlight to show a manageable local context for a seamless exploration.

The main interface to start explorations is the ③ Main Graph UI (see also Ⓒ in Fig. 1). It provides a highly scalable *GPU-based rendering* of a *Knowledge Graph* (a network-based visualization of the interconnected data items and their relationships), together with several linked views. This graph-based overview is less scalable than a matrix-based approach [14], however, aligns more closely with the mental image of analysts when exploring a network, as link charts have been used in investigative work for a long time [63]. The user can navigate this graph with a mouse and keyboard, select, hover, move, and (context) click individual nodes (data or extracted information items) and edges (their relations). The graph uses a selectable 2D or 3D node-link representation and is rendered using a force-directed layout. Strengths are calculated using centralizing, link, and charge forces based on a Barnes–Hut approximation. While this graph is initially automatically generated, the expert can (and is expected to) explore, interact, add, modify, and enhance it while working on the case. When modifying or judging information, user confidence in relationships (edges) can be encoded using the 6x6 intelligence scoring system for ③ relevance grading. The default confidence is F (Unknown), and for automated decisions that have not been manually reviewed, never above C (fairly reliable) to prevent Automated Inequality (R3) and wrong conclusions. All the interactions happen within the graph view or via individual visualization modules, which are reachable via the registered context actions and context menus, allowing for seamless transitions, which are appreciated by users.

The visual interface to the graph model has several features to enable Customization (A6) and User Agency (A1): A *sidebar* on the right offers several features: (1) control options for visualization (e.g., color, line thickness), layouting (force-direction layout strengths), and modes (e.g., 2D/3D-dimensionality, display modes for exploration like only displaying cross-matches, i.e., results from multiple documents) allow for customization and task-specific adaption, (2) an interactive *search* functionality allows filtering the graph quickly, (3) a *context display* shows information about a selected entity, and—leveraging the integrated graph model—occurrences, e.g., in text documents, (4) an overview of all available nodes and edges grouped by types, (de-)selectable individually or in groups.

A *timeline* at the bottom shows both datetime information as part of the Knowledge Graph and document times, allowing for brushing and filtering to optimize the graph view and empower investigators to follow an event- and time-based workflow in alignment with their exploration. When hovering over documents, only these are shown, while selecting zoomable and shiftable ranges restricts the shown parts of the graph. As can be seen from the examples, the amount of information displayed in the graph view is typically quite large, which hampers exploration. Therefore, a ② *Neighborhood Exploration*, acting similar to a magnifying glass or spotlight, allows to show the local neighborhood of a node (for example, 3 or 4 steps), and clicking any visible node transitions to the new neighborhood, allowing for a seamless exploration with a manageable amount of local, contextual information displayed without overloading the users, which can improving Efficiency (A4). Another approach to reducing the amount of clutter is to selectively merge confirmed relations to clusters, for example, aliases for persons or create groups. A slider allows for a confidence level based on the 6x6 system, which means that automated decisions without manual verification are never categorized as (very) likely (A or B), preventing Automated Inequality (R3) and enforcing Fairness (A3) and critical reflection (R4) through Human Oversight (R5).

Due to the amount of information shown (for typical investigations, this can be 30k nodes and 100k edges), we need to use several techniques to achieve 60+ fps performance: The graph is rendered entirely on the GPU and leverages instancing and custom shaders. This results in, once set up, a fixed-sized geometry of a few hundredth vertices and three WebGL draw calls (nodes, edges, labels), resulting in efficient rendering performance. Much of the visualization and visibility status is controlled from within the shaders, with crafted texture atlases and mipmapping for efficient textures, especially for nodes and text labels. To render more than 0.5 million text characters in real-time, we use a pre-generated font texture atlas and supply each node label instance with its correct, fixed-size ASCII-Code label (Unicode would be possible, but increase the texture size). This supply of instance-specific data (e.g., labels, position, node render state) is achieved through uniform buffer objects, acting similarly to a memory map, which is highly efficient. The sidebar uses virtual lists to render on demand, further reducing DOM usage. However, the number of nodes and edges is still limited by JavaScript and Browser performance.

The NER module offers an Ⓐ ontological search and Ⓑ textual view (see Fig. 1) as UI components. In the UI, a ① **context overlay** can be shown, for example, over a person's name with a preview image of a person together with other meta-data. This reduces domain boundaries and relieves the mental load of the users. Text understanding can be helped (see Section V-E) by color-coding named entities according to type and offering aggregation and interactions. Linked views at the bottom show all entities in the document grouped by type and ordering, e.g., by count, can be used to quickly navigate between occurrences through auto-scrolling, highlighting, and stepping.

The **ontological search** uses multiple (de-) selectable semantic search modes (exact match, substring match, fuzzy match, or ontological match). The latter allows searching *semantically* instead of guessing the correct keywords. This ontological search is considered very beneficial by the experts, as it reduces the burden on them to know the exact terms used but more generically describes the concept of what they are looking for. Search results are shown with specific probability scoring based on the distance (steps) taken in an ontology database, linking different properties. One example would be to search for ”accommodation” and get results with ”hut”, ”hotel”, or ”cottage”. The quality of the results, of course, depends on the extensiveness of the ontology, which often has to be adapted domain-specifically. Here, the experts can modify the ontology *on the fly*, e.g., to adapt to specific codewords.

Another type of interaction resulting from the tight integration comes even closer to the traditional visual analytics loop: While updating analysis parameters within a module usually only affects this module's results, through the fully-integrated data model and module listeners, it becomes possible to achieve *inter*-module exploration and refinement, coming closer to the expected levels of automation by current users. For example, when several speakers in audio files are recognized, and the transcripts are polluted by some of the speakers being background noise, the user can manually deselect the speakers, resynthesizing the audio, and the downstream analysis is automatically re-run, i.e., transcription and then knowledge extraction through NER. Old results can, in this process, be hidden (i.e., flagging the old document and its inference) to avoid an over-cluttering of the graph, which is considered extremely relevant by the experts to allow them to focus on relevant information only but can also be used to preserve Privacy (A2, C2).

While users work with the application, all performed actions are logged to achieve provenance, provide Accountability (C6), as well as prevent abuse through Human Oversight (R5).

V. Evaluation

We conducted a thorough evaluation of our approach, including feedback from multiple perspectives, to determine the effectiveness of the system. To showcase the practical usefulness of our approach, we present a case study in an investigative journalism setting, supporting war crime investigations (see Section V-A). To scrutinize the ethical and privacy risks involved, we then evaluate our approach based on ethics design guidelines [9] for intelligence applications (see Section V-B). To judge the resulting capabilities of the developed framework, we use a state-of-the-art intelligence capability assessment [15] (see Section V-C). To assess the quality of the underlying language model, we performed benchmarks on relevant NER-task, achieving state-of-the-art performance (see Section V-D). Finally, to evaluate the system from an expert perspective, we conducted a formative user evaluation with eleven domain experts in law enforcement (see Section V-E).

A. Case Study

In the following, we describe a simplified, *artificial* case study modeled after real-world workflows seen in **investigative journalism**. Here, we describe the process of identifying, placing, attributing, and documenting **war crimes**. We have chosen this example due to its high relevance, the high analysis stakes both for the victims as well as innocent persons, and the plausible availability of large amounts of multimodal data.

Goal — Alisa is an aspiring journalist for the respected newspaper *The Custodian*. She has been reporting about a brutal war in her home country for months now. While there have been some high-profile reports on war atrocities, she knows this is just the tip of the iceberg, and many people are missing. After reading some OSINT (Open Source Intelligence) reports, she wonders if she can also find out more about the forgotten victims of war. Simultaneously, she wants to see the perpetrators held accountable, so she aims to document her findings meticulously and hand her chain of evidence over to the ICC (International Criminal Court), which has started pre-trial investigations.

Data Collection — She starts off by collecting raw data: From various online sources reporting about the war, like Telegram, she exports messages, images, audio, and videos. From a friend and contact working for a large telecommunication provider, she gets a large dump of telephone calls and texts originating from foreign cell phone numbers logged into the telco's network. They were recorded by order of the nation's domestic intelligence agency. Further, on her newspaper's website, she allows for a SecureDrop submission for images and videos. Overall, she ⑤ receives thousands of hours of audio and video and tens of thousands of texts and images, which she imports into MULTI-CASE. The system ingests this data and runs the analysis pipeline.

Initial Exploration — First, Alisa is overwhelmed by the sheer amount of data in the Ⓒ *Knowledge Graph* view. She looks around and randomly starts listening to some recorded phone calls via the ① preview hover menu. Some are hard to understand due to multiple persons talking intermittently in the background.

Analysis Pipeline — The system offers her ① automatically-generated transcripts through the *Speech to Text* module while the audio is played simultaneously. She notices that the transcripts are not perfect when hearing the recordings, but they still help her a lot, as she can skim over the content in the Ⓑ *Document Viewer* much quicker. Wondering if the speakers talked about locations, she searched manually for common city names, finding many results. She realizes she can also use the entity search to display all locations the semantic text analysis has found, a summary of which is shown at the bottom. Through these and reading some context, she realizes the transcripts are intermingled with speech fragments (and locations) from the background speakers.

Multimodal Combinations — She ⑤ jumps back to the graph view and selects the *Speaker Recognition* module for the selected node. It identified four speakers and offered some best shots to listen to, together with individual transcripts. Hearing them in isolation, she realized that two were radio moderators. She deselects both speakers and lets the downstream analysis task run again. In the Ⓒ *Knowledge Graph* view, the old entry is ④ transparently archived and replaced by the new audio. Now, the recordings and transcripts are much clearer, but listening or reading through only a few would still take hours.

Semantic Search — She decides to Ⓐ *search* literally for some terms and words she suspects might have been used but get fewer results. Instead, she enables the fuzzy as well as the *ontological search*. Now she receives many more results. In some, the spelling seems off, and in others, she gets synonyms and hyponyms for her query. Reading over some of the matching sentences, she realizes several specific words are used and also learns some new ones the system did not detect.

Retraining On-The-Fly — She adds those words to the ④ *built-in ontology* and re-runs the search. As she reads a conversation about a small village where ”a lot of —— things happened,” she feels she might be on to something. Semantically searching through the remaining transcript in the Ⓑ *Document Viewer*, the speakers refrain from mentioning the village or such events again.

Cross-Matches — However, the system has recognized the village's name as a location descriptor and offers her to view it in the Ⓒ *Knowledge Graph* view. There, she uses the ② *Neighborhood Exploration* to see all connected entities up to three steps from this town. She finds out that another document mentions this tiny village in a spatial context to a larger town while the village is again allegedly mentioned in connection with some persons named A and B repeatedly over an extended time period. Using the Ⓒ *timeline view*, she restricts the view to a specific time range where she knows that the area around this larger city was temporarily invaded before the attackers were forced out into the neighboring woods. The graph becomes less crowded, and the system displays a weak link from person A to another name A‘ with a longer name form. The weak link comes from yet another transcript, where the persons named are mentioned closely together.

Manual Investigations — Alisa requests her assistant to read the transcript while she briefs her boss about the preliminary findings. After returning, Alisa sees that her assistant (working collaboratively on the case with her) has concluded that the persons mentioned in the report are likely similar and ③ has assigned a B score (highly likely) for the link in the 6x6 system [16]. The person A‘ has also been mentioned in the caption of a Telegram picture. Having used the *Image Analyzer* module, her assistant has found visual matches for this person in several pictures and also two videos, which he has flagged for her. She watches both videos, and one clearly shows a war crime.

Handling Fakes — She also Ⓐ searches for B, and she finds a graphic image but also sees B in a similar setting, seemingly taken weeks prior. She identifies the environment and obtains a broader view of the situation: the image is fake, likely disinformation. She ⑥ adds a comment and marks it as disproved, becoming archived by default.

Progressive Analytics — During her background research, interviewing one ICC representative, she is offered access to the ICC evidence collection platform, where users worldwide can upload materials of suspected war crimes. She also ⑥ imports this potential evidence enriching the underlying data model. Now, she runs a further person search using *Image and Video Analyzer* and finds the picture of a military photo ID. The person in the picture looks very similar to A.

Evidence Collection — Using the ⑪ reporting functionality, she prints out a trace of her analysis steps, including the transcripts with reference to the original audio files, the connection network with locations, all the associated imagery and data as a PDF report, and the associated document dump. She plans to hand it over to her ICC contacts and lawyers for them to further verify the potential claims for a subsequent trial. They plan to perform classical investigative work like forensic audio, facial analysis, and site visitation to collect evidence to back up and corroborate the potential war crime she found using the system, now knowing what to look out for.

B. Ethics Design Guidelines

In previous work [9], we have discussed in depth the ethical implications of using VA systems in intelligence and derived the first comprehensive overview of *detailed, technical* considerations to take into account when designing such systems. As pointed out, *the ethical implications [have to be considered] as an integral part of the design process from the outset* [9]. In the following, we describe how we have applied those considerations during the development of MULTI-CASE. The guideline numbers (e.g., C1-6, R1-5, A1-6) refer to the nomenclature established in previous work [9].

Semi-automated analyses are used, but the user remains in control for Human Oversight (R5), and the automated decisions are transparent (e.g., through ④ attribution and ③ confidence scoring) for Opacity (C3), addressing User Agency (A1) and Lack of Accountability (R1). ⑨ Provenance of the analysis steps taken can further strengthen this Human Oversight (R5) and provide Accountability (C6). The ability, for example, to ⑥ flag wrong or unrelated content can support Privacy (A2, C2) aspects by being less intrusive than human verification (as humans might memorize sensitive information). All automated system risk exhibiting inherent Discriminatory Bias (C1), but human operators also do. We published our underlying model for transparency reasons (cf. Opacity (C3)) and to detect or Prevent Automated Inequality (R3). The design as a hybrid Human-Machine-Configurations (C5) inherently ⑤ allows for mutual checks and balances to facilitate more Fairness (A3) and Human Oversight (R5). The semi-automated analysis undoubtedly can ② improve Efficiency (A4), while care was taken not to abstract too much and for the information to remain ① transparently attributable (cf. Opacity (C3), Accountability (C6), Lack of Accountability (R1)), which is achieved through the unified ④ fully-integrated data model. By making clear what aspects are automated and which are manual, by providing ③ confidence scores, and by not offering unrealistic features such as ”solve investigation” buttons, one works against Exaggerated Expectation (C4). Effective usage of the system and Literacy (A5) can only come with experience and daily usage. Integrated ⑥ sharing between colleagues, e.g., of saved search filters or information through comments, can support this. However, we note that more could be done here for our approach, but we expect that literacy will primarily be achieved through classical Training and Community-Building Among Users (R2). By enabling ④ interactive modifications to the underlying models like the ontologies, Customization (A6) can help users to adapt the system to their needs. One aspect to further improve upon is automated guidance to facilitate critical reflection (R4), for example, by automatically trying to detect biased behavior by human operators.

C. Intelligence Capability Assessment

We assess our framework according to a system capabilities classification [15]. This generic classification aims at knowledge exploration systems, including holistic approaches, focusing on intelligence applications. The classification's main focus is to assess the (technical) capabilities in a structured form, for example, if time-series data is supported, what type of interactions are used, or which type of knowledge is generated through AI support. In this regard, it indirectly includes results from older requirements studies in intelligence [41], [42], [43], [44]. However, these previous studies primarily describe the user interactions with the system like Jigsaw [46] through Overview and Detail, or Find the Clue and Follow the Trial [42]; those aspects included in the older studies but not in the capability assessment were evaluated as part of the expert evaluation (see Section V-E). We describe and assess our approach according to the 52 criteria posed in the **classification scheme** [15]. The icons indicate ○ no, ◐ partial, and ● full support. For a detailed discussion on the attributes themselves, we refer to the original paper while we provide examples and placement of MULTI-CASE's capabilities in the following:

In the dimension *Data and Information*, MULTI-CASE can compete with the state-of-the-art: It supports all basic **Data types**: *text* like documents and messages, *audio* like recordings, *image/video* like pictures or video recordings, *network* like relationship networks or call records, and ◐ *time-series*, primarily through meta-data like discrete timestamps. Classical, continuous time series are not explicitly supported. Regarding the **coding** of data, only *digital* modalities (i.e., the face-value of information) are supported, not ○ *analogical* ones (e.g., interpretation of facial expressions to detect lies or irony). This is comparable to the vast majority of approaches. Regarding the orthogonal **Expression**, *explicit* information is supported, but also *implicit* one, through the use of the underlying ontologies, which is a rare capability. With regards to communication between **Parties**, ▦ group communications are supported, but not specifically nested groups (i.e., subgroups). Analysis of ○ **Power Relations** is not supported. However, the investigative application is designed in such a way that it accounts for acts of deception and partially considers the ◐ **Measurement Problem**: For example, the use of code words is, in principle, supported through the domain-specific ontologies and specially trained NER model and also by looking at meta-data, which is harder to craft. This is a crucial capability in investigative systems, which many current approaches still delegate fully to the users.

In the dimension *Processing and Models*, our approach is suitable for a wide variety of analyses. Regarding the **Methodologies**, supported are *Representational* analysis to present the information, and especially *Confirmatory* analysis to validate hypothesis as well as *Exploratory* analysis to find relevant, a priori unknown facts. ◐ *Predictive* Analytics is partly integrated, depending on the employed modules. In terms of the employed analytical **Modalities**, all primary ones are equally supported: *Content* like actual text or videos, for example, through the Document Viewer or the Video Analyzer, *Network* for relationship analysis through the Knowledge Graph and Neighborhood Exploration, or *Meta-Data* through the Knowledge Graph and the Timeline in combination with the filtering functions. This holistic, integrated, and interconnected analysis is a crucial factor distinguishing MULTI-CASE from most existing approaches. The **Analysis** itself supports an incremental, streamed data import, making it an *online* analysis. Regarding the **Latency**, the standard use case for an investigative system will be a *delayed* 🅳 analysis. One key advantage of the underlying model and the modular architecture is the achieved **Scalability**. It supports huge (IIII) investigative volumes for *ingress*. Also, through its Neighborhood Exploration, the number of concurrent entries under consideration in the *analysis* can be regarded as medium (II), more than many other approaches. As our approach is a research prototype and not a commercial application, the support for ○ **Data-Mappings**, like many importers, is limited.

In the dimension *Visual Interface*, many combined strategies are leveraged. Regarding the visualization **Pane**, the usual *2D* is supported, but the Knowledge Graph also leverages *3D*. Stereoscopic 3D ○ *S3D* is unsupported but easily addable. Regarding the **Operation Methods** [64], all are supported: one can *Select* an entity to get more detailed information from all modules combined, *Explore* different semantic matches or the Knowledge Graph, *Reconfigure* the confidence thresholds for automated merging, *Encode* the data as inferred graph relation representation, *Abstract/Elaborate* by adapting the Neighborhood level or inspect information within a specialized module, *Filter* trough the semantic search or a timeline range, and *Connect* by showing graph neighborhoods. The **Manipulation** happens both *directly*, e.g., by selecting entries, and *indirectly*, for example, by choosing specific analysis modes, for example, only showing corroborated cross-matches. The **Goal** of the actions is primarily *data tuning* to show relevant information. However, the approach also partly supports ◐ *model tuning*, where the interactions influence an underlying mode, e.g., by manually confirming relationships between entities or updating the ontology. The **Strategy** involved in interactions are both *iterative* and *progressive*, which go hand in hand in investigative scenarios. The ◐ *active learning* depends on the individual analysis modules, through feedback or showing an example.

In the dimension *Knowledge Generation*, the **Explanation** of information is performed through *numerical*, *textual*, and *graphical* representations, for example through scoring/sorting, annotated highlighting, or charting, respectively. The **Transfer Function** operates both on the *machine model*, which updates the underlying model through interactions, as well as the *mental model* of the analysts. **Factors** that are considered in our approach are *confidence*, *trust* and *privacy*. For a more detailed discussion, see Section V-B. With regards to the **Time Dimensionality** of the Knowledge Generation, the approach primarily enables the exploration of past information but also allows conclusions based on this information for the given dataset ▦ . The **Predictive Power** of the system relates to explaining past events and potential upcoming links but also forms predictions about yet-to-ingest data ▦ . Regarding the **Evaluations** performed, we present a case study 🅔 (see Section V-A), this capability assessment ↔ (see Sections V-C-V-B), as well as an expert evaluation 👤 (see Section V-E).

D. Model Evaluation

We evaluated *our NER model* (Huggingface model via osf.io/eap4r) and five *baseline models* based on a hold-out test set of the re-tagged news dataset that we publish for future benchmarks. Based on preliminary experimental results, we decided to train our own NER classifier based on the weights of the pre-trained GottBERT [17] language model. We report precision, recall, and F1-score for each entity label (see Tab. I). As an overall observation, we find that our test dataset is challenging for the NER model, as the achieved performance is below scores reported for existing benchmark datasets like GermEval2014 [56] for all models, including the baselines.

The best-reported score for the GermEval dataset is 86.8% [17] with categories *PERSON*, *LOCATION*, *ORGANISATION* and *MISC*. However, on both our generic news dataset and, in particular, the scenario-specific text data, we see significantly lower performance. Still, our model outperforms or matches baseline models on the core categories while achieving satisfactory performance on most additional categories. Still, we observe a drop in performance in broad, newly introduced categories like *EVENT* or *PRODUCT*.

E. Domain Expert Evaluation

To showcase the effectiveness of our approach in comparison to existing methods, we conducted an expert evaluation with eleven domain experts (LEA 1-2, RS 1-3, SI 1-3, LAW 1, EE 1, POL 1) working in the context of law enforcement.

Expertise LEA 1 is a recently retired former special police forces commander with a 40-year career, now working as a security consultant for law enforcement agencies. LEA 2 is a leading investigator at a federal police force with a 20-year career investigating organized crime. RS 1 is a research scientist and head of the research department with a 30-year career in speech recognition. RS 2 is a junior researcher and developer working for a federal security agency developing analytical solutions for law enforcement in the area of digital forensics. RS 3 is a junior researcher working for the same federal security agency on the topic of phone analysis. SI 1 is a senior principal research engineer with an almost 30-year career overseeing numerous identity solution projects for an

TABLE I
VALIDATION ACCURACY FOR THE FIVE BASELINES (DE_CORE_NEWS_{SM,MD,LG}, BERT-GER, XML-ROBERTA) AND *our* NER MODEL

	sm			md			lg			BERT-German			XML-RoBERTa			**Ours**		
Type	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PERSON	.69	.72	.70	.76	.77	.77	.78	.78	.78	.93	**.89**	**.91**	.91	.87	.89	**.94**	.88	**.91**
ORGANIZATION	.55	.47	.51	.56	.52	.54	.59	.55	.57	.81	.65	.72	.75	.64	.69	**.78**	**.78**	**.78**
LOCATION	.53	.57	.54	.59	.61	.60	.61	.61	.61	**.90**	.63	.74	.84	.62	.71	.88	**.90**	**.89**
MISC (Original)	.14	.29	.19	.17	.37	.24	.18	.36	.24	-	-	-	**.30**	**.45**	**.36**	-	-	-
MISC (Own)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.15	.22	.18
EVENT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.99	.40	.57
PRODUCT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.49	.59	.54
DATETIME	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.99	.99	.99
LANGUAGE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.98	.95	.96
LAW	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.60	.60	.60
QUANTITY	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.97	.96	.97
NUMBERS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.98	.98	.98

international security company. SI 2 is a project manager with a 25-year career working on video analysis and investigative systems at the same company. SI 3 is a principal research scientist with more than ten years of experience in video object tracking also at this company. LAW 1 is a professor and criminologist specializing in security management, hate crimes, and legal aspects with more than 15 years of experience in the field. EE 1 is a sociologist and ethics advisor offering guidance for security research projects. POL 1 is a project supervisor at a national project management agency overseeing civil security research and policy expert.

Methodology The expert evaluation was conducted as a formative evaluation and took a combined 180 minutes, split into a 60-minute presentation and a 120-minute hands-on evaluation. The 60-minute introduction delivered to all experts described the capabilities of the system on a conceptual level while also demonstrating actions in the form of one to three-minute-long screen recordings. During the evaluation, a single station (27-inch FHD screen, mouse, and keyboard) with the prototype was available to the experts, together with two researchers standing by to help with questions and advice. During this time, one of the experts would typically use the system to explore the prototype while being encouraged to think aloud. The other experts could meanwhile observe, comment, and ask questions. After irregular intervals, the experts switched positions, and usage time between experts varied between five to 20 minutes. During the whole session, the experts were asked questions aligned with a semi-structured interview sheet containing a set of 38 prepared questions covering various aspects of our approach. The session's aim was to elicit the domain experts' opinions about the system and gain insights into how they would use the system in their investigative workflows. Further, the experts were asked to comment on the approaches' capabilities, user-interaction concepts, and visualizations while identifying opportunities for improvements. The detailed findings of this evaluation are presented in the following.

Findings Asked about the **benefits** they see in an investigative framework like MULTI-CASE, the criminal investigators state that they hoped to be relieved of the time-consuming, "extremely high manual workload, which currently requires much personnel" (LEA 1) "and time" (LEA 2). Of course, there are existing use case management systems, but "their usage and the casework is performed very much in a manual way [. . .] with little technical support" (LEA 1), which becomes a "big problem for mass data" (LEA 2), where "automation can be very helpful" (RS 2). In "particular observations produce very large amounts of video data" (LEA 1). For particular problems, some isolated technical solutions exist at some local partners, for example, geo-based analyses (cf. LEA 1), but access depends on the local support and willingness of the partners to help (cf. LEA 1). Further, one of the most important features for them is to import many different types of multimodal documents like "existing records, images, videos" (LEA 2). Here, MULTI-CASE as a ❶ "large overview system for multimodal data like audio, text or video has the potential to drastically improve investigative work" (RS 3), making it "uncharted territory" (LEA 1). The other experts strongly agree, noting that currently they lack "a complete picture [in a single system]" (LEA 1) and "nothing in this form exists" (LEA 1): neither for phones (cf. RS 3), speech (cf. LEA 2), or text (cf. LEA 1). "Multimodality is the largest benefit, as everything can be seen in context" (RS 2).

Regarding the **risk of automation**, they are aware of potential pitfalls but do not consider them highly problematic: It is likely that "there are errors in the analysis" (RS 1), for example, by different spellings (cf. RS 1). This, however, can also happen when case workers need "to read through thousands of pages or watch weeks of video recordings, where things might be overlooked and error rates increase with time as frustration increases" (cf. LEA 1). "From an automated perspective, it might not be most important to find everything, but to start and find many relevant things" (POL 1). From a "legal perspective, this might be much more critical, as innocent individuals can become part of an investigation" (RS 1). They note that "automated analysis is less of a problem when there is reasonable suspicion for a suspect, but an infringement on fundamental rights is" (LAW 1). In this regard, the modality differs: "images are considered more critical than voice, which in turn is more critical than text" (cf. LAW 1). Possible ways to solve this are "by not focusing on the subject, but on the right infringements [for involved parties]" (cf. LAW 1). This means automated analysis has the potential to be considered less invasive than manual analysis, but "for example through data economy and short-term storage, but this depends on the case" (EE 1).

From an **ethical perspective**, it might be more "justifiable to let the computer search for targets instead of humans" (EE 1) as the human "remembers" (EE 1) offering potential for misuse, while the system forgets after the comparison. Current approaches "do not consider privacy or ethical aspects sufficiently" (LEA 2) and the investigators are independently responsible on their own to follow the rules - however, "there is a wide gap between theory and practice" (LEA 2). "A verified system that works with high accuracy [and without bias] could be fairer than an arbitrary human" (EE 1), as "many humans are very selective and inherently biased" (LEA 1). Regarding the fear of intransparent, autonomous decisions, it was noted that "the systems are always support systems and humans always the final instance" (LEA 1), and before a "prosecution will always be manually verified" (LEA 1). A problem can arise when "humans become too careless and trust the system too much" (EE 1).

One interesting discussion arose regarding the error rate: From the perspective of an analyst "false positives are less of an issue as they can be manually verified, while false negatives are missed" (LEA 1). From the "perspective of innocents, this is directly inverse, but this again depends on the context" (cf. EE 1). "When misses lead to extreme dangers for others, this can be very bad" (cf. EE 1).

The experts consider ⑥ **collaboration** features relevant, where multiple users can work on the same case, as they sometimes have to work with "widely distributed experts" (cf. LEA 1). Also, the "parallel work between colleagues is nice" (EE 1).

Regarding the central **Knowledge Graph**, many experts agree that it can provide a key overview, as "it is extremely important to show all the relations" (POL 1) and the "connections" (LEA 1), which is a "large advantage" (RS 2). "Showing everything together is very relevant for keeping an overview" (SI 1). For this, the ② *Neighborhood Exploration* is considered "a must, especially when many data items are loaded" (RS 2), as it allows to reduce the visual clutter and only show contextual information. This is an example of a filtering functionality, which is regarded as "essential" (RS 2). Also, the ability to filter the graph and the mergings by ③ confidence is regarded to be beneficial (RS 1). Similarly, the timeline is also considered "very helpful" (RS 2), as "the time and event sequence is very important for the investigation" (LEA 2). In this context, the interactions are regarded as "very smooth and nice looking" (RS 2). However, some experts questioned "if 3D is necessary" (cf. LEA 1) and would favor the 2D graph that is also available. The graph view can act as a "supportive mental map [. . .] and a large digital notebook" (LEA 2), which "currently is often only in ones head" (LEA 2). For this, the "comment function" is essential and helpful (cf. LEA 2 and EE 1) to make notes, which can be shared between users. Regarding the confirmatory investigative work, however, the "graph view is less important" (LEA 1), where the "individual analysis modules like the document viewer or audio analysis are more relevant [. . .] supporting the daily work" (LEA 1). For example, in the document viewer, the "automated recognition of entities in the document which are shown at the bottom with their number of occurrences, is especially helpful, as it allows to get a ④ summary understanding of the content of the text already" (cf. LEA 1). Also, the automated transcription of audio "given sufficient quality, is very important and a key advantage" (cf. LEA 1). Especially relevant is the ability to seamlessly switch between view and modalities, for example, ⑤ "to jump from a node in the graph to the text location in the document viewer" (LEA 1) as well as "jumping to search matches" (LEA 1). However, it was noted by several experts that a proficient usage would "require training" (cf. LEA 1, EE 1, RS 1, LAW 1), after its completion, however, would be a "productivity boost" (EE 1).

In terms of **potential future features**, some ideas were mentioned: Among expected quality-of-life improvements like more file type support (cf. RS 2), one area of improvement could be group conversations (cf. RS 2), for example, through colored attributions also inside the document viewer, the creation of cluster-nodes in the graph view to merge related, but currently less interesting entities (cf. RS 2) or show a modification and usage history from co-workers (cf. RS 3). Also, for the comments and exploration history of colleagues, a "misuse button" (cf. EE 1) would potentially be useful to report incorrect use. Also, some more explainability for the automated parts, i.e., why a "speaker was recognized" (cf. EE 1) as such, would be useful and increase trust. Overall, the approach "will be well usable for semi-automated investigative analysis [. . .] between a knowledgeable user and a supportive system" (LEA 1).

VI. Discussion and Future Work

As we demonstrated, our approach enhances the capabilities for multimodal intelligence analytics. In the following, we discuss the valuable lessons we learned during development, the implications of the valuable feedback we received about our prototype, architectural design trade-offs, limitations of the approach, and potential future work that remains.

A. *Findings and Lessons Learned*

Based on the evaluations in the previous section, we have succeeded in working towards fulfilling the experts' requirements posed from the beginning: MULTI-CASE is an exemplary centralized, multimodal platform framework that allows several analysts to collaboratively work on cases and empowers users through the transparent inclusion of AI-aided decision-making while relieving them of burdensome tasks and considering ethical design guidelines. Following the UNODC [16] task definitions, the main tasks can be performed: link analysis between entities is supported while also allowing to consider them in the context of the surrounding events based on a timeline. While it supports a basic flow analysis in principle, the visualization modules presented here are not particularly suited for this analysis, but through the modular design, a component operating on the shared data model could be developed. We have seen how the multimodal approach can support the analysis of otherwise difficult-to-detect cross-matches, while a visual analytics-based approach has benefits in terms of agency, accountability, and trust. The experts are open to AI-based solutions, especially when it relieves them of mundane tasks, and they feel supported. Leveraging both computational power and human intuition in a tight feedback loop can positively influence the capabilities of the resulting human-machine configuration. Regarding the displayed results, they tend to believe them at face value to some degree when they seem plausible, somewhat similar to findings reported to them by colleagues.

However, we also saw that experts have high expectations regarding the machine results and—especially when not specifically trained for the system—are rather unforgiving with respect to unexpected or contradictory results. Also, they can be easily annoyed in case they feel the system hinders them, holds them back, or torments them through seemingly obvious confirmations. Based on these observations, we can derive several key findings:

F1: A Holistic Approach Supports Finding Cross-Matches
The case study and expert evaluation shows that intelligence investigations require interconnected, multimodal analytics.
Implication: A holistic approach can combine these different analysis modalities within a single context, reducing domain boundaries and enabling effective search for cross-matches. Especially relevant here is a vertical integration between all analysis modules, for example, through a fully-integrated data model.

F2: Unobtrusive Support-Systems are Accepted
As long as a system remains supportive and unobtrusive, relieving analysts of mundane tasks or providing them with valuable hints and insights on request or through nudging, semi-automated systems are accepted. Tormenting approaches hindering the workflow or being intuitive or unreliable can destroy an initial level of trust placed in the system.
Implication: A self-explanatory, easy-to-use user interface combined with helpful but unobtrusive functions is essential. For this, the right balance has to be found between automation and manual confirmation. Unreliable or inconsistent results (without indications) or hindering of workflows should be strongly avoided

F3: Reduce False Negatives for VA—and False Positives for AI
Initially surprising to us was that for many tasks (e.g., search, filter, linking), the domain experts (both LEA 1-2 and EE 1) prefer the error rate to depend on the automation level: for semi-interactive VA a reduction of false-negatives is often preferable, while automated systems should reduce false-positives.
Implication: Consider the optimization task carefully, as the *cost of error*, where not finding something (i.e., FN) or a wrong attribution (i.e., FP) is considered more costly than the opposite. A missed lead might break the whole investigation, while a wrong attribution might cause serious harm to innocents.

F4: Limited Acceptance of Unreasoned Decisions
At least for now, to support an ethical and privacy-aware analysis and offer transparency, fairness, and accountability while fostering user trust, the experts prefer an explainable, interactive system compared to a fully automated approach.
Implication: Due to the high stakes in this domain, experts have concerns about fully-automated systems that cannot provide a rigorous chain of evidence, which—at least for now—is rarely possible. Future developments might shift this balance.

B. Limitations and Future Work

Nevertheless, the approach remains a research project with limitations:

The **Knowledge Graph** representation uses custom GPU-optimized rendering achieving excellent performance, but it comes with the disadvantage that some of the more advanced results from graph drawing, like more complex curved lines, are not directly applicable without heavy performance penalties. We also want to highlight that we do not see our contribution in designing state-of-the-art graph drawing but in the interactions, combinations, and linkings between the different modalities for the graph.

The integration of the **underlying language model** itself is modular, such that any other transformer-based NER model can be easily used, as the system features a built-in language detection. However, for the evaluation in this chapter, we only used our customized German NER model due to the domain experts' preferences and expertise. We did not explicitly show a generalization, which we, nevertheless, certainly expect. In the future, off-the-shelf transformer-based NER models can be used, with limitations in the types of detected NER and resulting degradation in relationship inference. Alternative models would need to be fine-tuned with additional NER types, requiring appropriate training data. Another problem in this regard can be the analysis of multi-lingual or inter-lingual text and transcripts.

The recent progress with **Large Language Models (LLMs)** like GPT-4 [65] offers interesting opportunities in this regard. This is, in particular, relevant when models are capable of supporting multiple languages as well as providing up-to-date and case-specific query context, as the *New Bings* underlying Prometheus Model [66] shows to some limited degree. Three domain experts (LEA 1, SI 1, SI 2) in our study tried Chat-GPT on crafted case material and were astonished both by the easy workflow of querying and the (relative) quality of the findings as potential leads. They regard such *text-based, interactive prompting* through LLMs, which imitates basic reasoning and summarization capabilities, as potentially very useful. Integrating such natural language prompts in applications, maybe only in supportive roles, seems very promising. Interestingly, GPT-4 also shows surprising capabilities in zero-shot NER labeling. For testing, we let GPT-4 auto-label a subset of our test data. We achieved this zero-shot labeling by prepending a prompt ”Extract named entities of the given types from the following text: person, organization and location”, resulting in only slightly less quality than manual, human labeling. This could potentially replace specifically trained NER models, like the one we described in Section III. Recent experiments [67] suggest superior results are possible. While this shows the viability of the transfer learning approach, ”close-to-real-life” scenarios often perform worse compared to controlled benchmarks [68]. Therefore, evaluating such scenarios in the wild is important to identify persisting limitations, which can be supported by interactive analysis. Also, care must be taken to consider the additional risks involved when using LLMs: They do not learn from mistakes outside their limited context window (32k for GPT-4-32k), which is relevant when using all documents as context, and most seriously, they tend to suffer from hallucinations that are hard to detect. Further, employment of such solutions would require on-premise solutions or specialized contracts.

Overall, depending on the jurisdictions, **legal requirements** might regulate the allowed automated analysis tasks [69]. The ethical and privacy-aware design, as well as the semi-automated analysis, always subject to human verification, performed in our approach, should allow for usage even in tightly regulated jurisdictions. The concrete usage in critical cases, however, should be accompanied by a prior legal counsel.

One general limitation in this line of research is the **opaqueness of the intelligence community**. Many systems are classified [21] and capabilities are not shared openly–which can be frustrating from a scientific perspective, hampering progress and introducing problems from ethical and privacy perspectives due to missing accountability. It also remains difficult to recruit domain experts to evaluate and analyze the techniques developed in the scientific community. One way to reduce increased reliance on expert evaluations is to also incorporate general interaction strategy design guidelines derived from numerous user interaction evaluations regarding relevance feedback [70]. Efforts are ongoing to finance more research in open and accountable intelligence solutions (e.g., within Horizon Europe and others). However, we are well aware that some aspects of this domain will likely remain hidden. With our work, we try to contribute to ongoing research in this domain and discuss ways to make these more accountable.

VII. Conclusion

Over the last few years, AI-driven models have become increasingly prevalent in many domains. This tendency can also be observed in operational analytics solutions in investigative journalism, intelligence, or law enforcement. These domains, in particular, pose distinct challenges due to their sensitive nature. Two aspects, in particular, stand out: ethical and privacy concerns, as well as difficulties in efficiently combining heterogeneous data sources for multimodal analytics. A lack of such *holistic and multimodal* approaches can lead to biased results and increased manual efforts through domain discontinuities.

To address these two challenges, we present MULTI-CASE, a holistic visual analytics framework that enables the exploration and assessment of heterogeneous information spaces (i.e., unstructured, diverse, and multimodal) supported by an equal joint agency between humans and AI to ensure ethics and privacy awareness. To fulfill these requirements, the system operates on a fully-integrated data model while featuring type-specific analyses with multiple linked components, including a modality-wide search (i.e., full-text, semantics, and all multimodal analysis results), text, and graph-based analysis. Different information streams are linked in a knowledge graph, providing in-situ explanations and transparent source attributions while facilitating responsible exploration through numerous interlinked explorative modules. We discuss the potential for improvements, for example, in rendering, completeness, or the use of more advanced LLMs.

We demonstrate how our framework fulfills the design goals through state-of-the-art intelligence capability assessments and evaluations according to ethics design guidelines. The underlying transformer model showed state-of-the-art performance on relevant benchmarks. To showcase our prototype’s analytical capabilities in practice, we presented a case study describing war crime investigations in the context of investigative journalism. Finally, a formative expert evaluation with eleven domain experts in law enforcement confirms that MULTI-CASE facilitates human agency and steering in security-sensitive, AI-supported analysis, addresses ethical and privacy concerns, and provides much-needed analytical capabilities.

With this contribution, we aim to provide more insights into the often opaque workings of the intelligence community and strive towards a more accountable and responsible use of modern AI capabilities.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) in the framework of PEGASUS and VIKING under the program ”Forschung für die zivile Sicherheit 2018 - 2023” and its announcement ”Zivile Sicherheit - Schutz vor organisierter Kriminalität II”.

References

[1] A. Mitra, *Modern Day Surveillance Ecosystem and Impacts on Privacy*. Hershey: Information Science Reference, 2021.

[2] J. Bullock, M. M. Young, and Y.-F. Wang, “Artificial Intelligence, Bureaucratic Form, and Discretion in Public Service,” *Information Polity*, vol. 25, no. 4, pp. 491–506, 2020.

[3] B. Ganor, “Artificial or Human: A New Era of Counterterrorism Intelligence?” *Studies in Conflict & Terrorism*, vol. 44, no. 7, pp. 605–624, 2021.

[4] M. Broussard, N. Diakopoulos, A. L. Guzman, R. Abebe, M. Dupagne, and C.-H. Chuan, “Artificial Intelligence and Journalism,” *Journalism & Mass Communication Quarterly*, vol. 96, no. 3, pp. 673–695, 2019.

[5] J. Stray, “Making Artificial Intelligence Work for Investigative Journalism,” *Digital Journalism*, vol. 7, no. 8, pp. 1076–1097, 2019.

[6] K. J. Hayward and M. M. Maas, “Artificial Intelligence and Crime: A Primer for Criminologists,” *Crime, Media, Culture: An International Journal*, vol. 17, no. 2, pp. 209–233, 2021.

[7] A. S. Obendiek and T. Seidl, "The (False) Promise of Solutionism: Ideational Business Power and the Construction of Epistemic Authority in Digital Security Governance," *Journal of European Public Policy*, pp. 1–25, 2023.

[8] D. Mügge, "The Securitization of the EU's Digital Tech Regulation," *Journal of European Public Policy*, pp. 1–16, 2023.

[9] M. T. Fischer, S. D. Hirsbrunner, W. Jentner, M. Miller, D. A. Keim, and P. Helm, "Promoting Ethical Awareness in Communication Analysis: Investigating Potentials and Limits of Visual Analytics for Intelligence Applications," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, 2022, pp. 877–889.

[10] B. Shneiderman, "Bridging the Gap Between Ethics and Practice," *ACM Transactions on Interactive Intelligent Systems*, vol. 10, no. 4, pp. 1–31, 2020.

[11] C. Rigano, "Using Artificial Intelligence to Address Criminal Justice Needs," *National Institute of Justice Journal*, vol. 280, no. 1-10, p. 17, 2019.

[12] P. M. Asaro, "AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care," *IEEE Technology and Society Magazine*, vol. 38, no. 2, pp. 40–53, 2019.

[13] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A Review of Predictive Policing from the Perspective of Fairness," *Artificial Intelligence and Law*, vol. 30, no. 1, pp. 1–17, 2022.

[14] M. T. Fischer, D. Seebacher, R. Sevastjanova, D. A. Keim, and M. El-Assady, "CommAID: Visual Analytics for Communication Analysis through Interactive Dynamics Modeling," *Computer Graphics Forum*, vol. 40, no. 3, pp. 25–36, 2021.

[15] M. T. Fischer, F. Dennig, D. Seebacher, D. A. Keim, and M. El-Assady, "Communication Analysis through Visual Analytics: Current Practices, Challenges, and New Frontiers," in *2022 IEEE Visualization in Data Science (VDS)*, 2022, pp. 6–16.

[16] UNODC, "Criminal Intelligence: Manual for Analysts," Vienna, Austria, 2011. [Online]. Available: https://www.unodc.org/documents/organized-crime/Law-Enforcement/Criminal_Intelligence_for_Analysts.pdf

[17] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine, and M. Boeker, "GottBERT: A Pure German Language Model," *arXiv preprint arXiv:2012.02110*, 2020.

[18] C. Groenewald, B. L. W. Wong, S. Attfield, P. Passmore, and N. Kodagoda, "How Analysts Think: How Do Criminal Intelligence Analysts Recognise and Manage Significant Information?" in *2017 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2017, pp. 47–53.

[19] W. Elm, S. Potter, J. Tittle, D. Woods, J. Grossman, and E. Patterson, "Finding Decision Support Requirements for Effective Intelligence Analysis Tools," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, no. 3, pp. 297–301, 2005.

[20] J. Scholtz, "Metrics for Evaluation of Software Technology to Support Intelligence Analysis," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, no. 10, pp. 918–921, 2005.

[21] M. K. Dhami, "A Survey of Intelligence Analysts' Perceptions of Analytic Tools," in *2017 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2017, pp. 131–134.

[22] D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[23] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[25] J. Y. Lee, F. Dernoncourt, and P. Szolovits, "Transfer Learning for Named-Entity Recognition with Neural Networks," *arXiv preprint arXiv:1705.06273*, 2017.

[26] S. Barbosa and S. Milan, "Do Not Harm in Private Chat Apps: Ethical Issues for Research on and with WhatsApp," *Westminster Papers in Communication and Culture*, vol. 14, no. 1, pp. 49–65, 2019.

[27] F. Sperrle, D. Ceneda, and M. El-Assady, "Lotse: A Practical Framework for Guidance in Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, 2022.

[28] A. Zytek, D. Liu, R. Vaithianathan, and K. Veeramachaneni, "Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1161–1171, 2022.

[29] M. Correll, "Ethical Dimensions of Visualization Research," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos, Eds. New York, NY, USA: ACM, 2019, pp. 1–13.

[30] N. A. Tu, T. Huynh-The, K.-S. Wong, M. F. Demirci, and Y.-K. Lee, "Toward Efficient and Intelligent Video Analytics with Visual Privacy Protection for Large-Scale Surveillance," *The Journal of Supercomputing*, vol. 77, no. 12, pp. 14 374–14 404, 2021.

[31] R. C. Geyer, T. Klein, and M. Nabi, "Differentially Private Federated Learning: A Client Level Perspective," 2017.

[32] P. G. Shynu, H. Md. Shayan., and C. L. Chowdhary, "A Fuzzy based Data Perturbation Technique for Privacy Preserved Data Mining," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2020, pp. 1–4.

[33] V. Batagelj and A. Mrvar, "Pajek - Analysis and Visualization of Large Networks," in *Graph Drawing*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, pp. 77–103.

[34] K. Schwarz and R. Creutzburg, "Design of Professional Laboratory Exercises for Effective State-of-the-Art OSINT Investigation Tools - Part 3: Maltego," *Electronic Imaging*, vol. 33, no. 3, pp. 45–1–45–23, 2021.

[35] H. A. D. E. Kodituwakku, A. Keller, and J. Gregor, "InSight2: A Modular Visual Analysis Platform for Network Situational Awareness in Large-Scale Networks," *Electronics*, vol. 9, no. 10, p. 1747, 2020.

[36] M. Dowling, N. Wycoff, B. Mayer, J. Wenskovitch, S. Leman, L. House, N. Polys, C. North, and P. Hauck, "Interactive Visual Analytics for Sensemaking with Big Text," *Big Data Research*, vol. 16, pp. 49–58, 2019.

[37] J. Zahalka and M. Worring, "Towards Interactive, Intelligent, and Integrated Multimedia Analytics," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2014, pp. 3–12.

[38] A. Wu and H. Qu, "Multimodal Analysis of Video Collections: Visual Exploration of Presentation Techniques in TED Talks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 7, pp. 2429–2442, 2020.

[39] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu, "EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 927–937, 2020.

[40] T. Tang, Y. Wu, Y. Wu, L. Yu, and Y. Li, "VideoModerator: A Risk-aware Framework for Multimodal Video Moderation in E-Commerce," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 846–856, 2022.

[41] J. Decker, A. Godwin, M. A. Livingston, and D. Royle, "A Scalable Architecture for Visual Data Exploration," in *IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST, J. Stasko and J. J. van Wijk, Eds. Piscataway, NJ, USA: IEEE, 2009, pp. 221–222.

[42] Y.-a. Kang, C. Gorg, and J. Stasko, "Evaluating Visual Analytics Systems for Investigative Analysis: Deriving Design Principles From a Case Study," in *IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST, J. Stasko and J. J. van Wijk, Eds. Piscataway, NJ, USA: IEEE, 2009, pp. 139–146.

[43] Y.-a. Kang and J. Stasko, "Characterizing the Intelligence Analysis Process: Informing Visual Analytics Design Through a Longitudinal Field Study," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2011, pp. 21–30.

[44] Y. Lu, R. Kruger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski, "Integrating Predictive Analytics and Social Media," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2014, pp. 193–202.

[45] T. A. Keahey and K. C. Cox, "VIM: A Framework for Intelligence Analysis," in *IEEE Symposium on Information Visualization*. IEEE, 2004, pp. p22–p22.

[46] J. Stasko, C. Gorg, Z. Liu, and K. Singhal, "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," in *2007 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 131–138.

[47] DataWalk Inc., "DataWalk," 2020. [Online]. Available: https://datawalk.com/

[48] Nuix Pty Ltd, "Nuix Discover and Nuix Investigate," 2020. [Online]. Available: https://www.nuix.com/products

[49] IBM, "i2 Analyst's Notebook," 2020. [Online]. Available: https://www.ibm.com/us-en/marketplace/analysts-notebook

[50] Palantir Technologies, Inc., “Gotham,” 2020. [Online]. Available: https://www.palantir.com/palantir-gotham/

[51] M. Scott, “How Ukraine used Russia’s digital playbook against the Kremlin,” *POLITICO*, 2022. [Online]. Available: https://www.politico.eu/article/ukraine-russia-digital-playbook-war/

[52] J. W. Mohr, R. Wagner-Pacifici, R. L. Breiger, and P. Bogdanov, “Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics,” *Poetics*, vol. 41, no. 6, pp. 670–700, 2013.

[53] L. Ratinov and D. Roth, “Design Challenges and Misconceptions in Named Entity Recognition,” in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ser. CoNLL ’09. USA: Association for Computational Linguistics, 2009, pp. 147–155.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” *Advances in neural information processing systems*, vol. 30, 2017.

[55] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew, “HuggingFace’s Transformers: State-of-the-art Natural Language Processing,” *CoRR*, vol. abs/1910.03771, 2019.

[56] D. Benikova, C. Biemann, M. Kisselew, and S. Pado, “Germeval 2014 Named Entity Recognition Shared Task: Companion Paper,” in *Workshop Proceedings of the 12th edition of the KONVENS conference*, 2014, pp. 104–112.

[57] D. Schwimmbeck, “HuggingFace: Domischwimmbeck/Bert-base-german-cased-fine-tuned-ner,” 2022. [Online]. Available: https://huggingface.co/domischwimmbeck/bert-base-german-cased-fine-tuned-ner

[58] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised Cross-Lingual Representation Learning at Scale,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

[59] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations (ICLR)*, 2015.

[60] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *International Conference on Learning Representations (ICLR)*, 2019.

[61] Jeremy Howard and Sebastian Ruder, “Fine-tuned Language Models for Text Classification,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 328–339, 2018.

[62] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” 2022.

[63] M. K. Sparrow, “The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects,” *Social Networks*, vol. 13, no. 3, pp. 251–274, 1991.

[64] J. S. Yi, Y. A. Kang, J. Stasko, and J. Jacko, “Toward a Deeper Understanding of the Role of Interaction in Information Visualization,” *Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[65] OpenAI, “GPT-4 Technical Report,” 2023.

[66] J. Ribas, “Building the New Bing,” 2023. [Online]. Available: https://www.linkedin.com/pulse/building-new-bing-jordi-ribas/?src=aff-ref&trk=aff-ir_progid.8005_partid.10078_sid._adid.449670

[67] F. Gilardi, M. Alizadeh, and M. Kubli, “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks,” 2023.

[68] A. Paleyes, R.-G. Urma, and N. D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2022.

[69] BVerfG, “Urteil des Ersten Senats (16.02.2023), BvR 1547/19, Rn. 1–178, ECLI:DE:BVerfG:2023:rs20230216.1bvr154719,” 2023. [Online]. Available: ECLI:DE:BVerfG:2023:rs20230216.1bvr154719

[70] O. S. Khan, B. Jónsson, J. Zahálka, S. Rudinac, and M. Worring, “Impact of Interaction Strategies on User Relevance Feedback,” in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, W.-H. Cheng, M. Kankanhalli, M. Wang, W.-T. Chu, J. Liu, and M. Worring, Eds. New York, NY, USA: ACM, 2021, pp. 590–598.