Forensic Science International: Digital Investigation 54 (2025) 301937

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

Autocrime - open multimodal platform for combating organized crime

Srikanth Madikeri [a,i], Petr Motlicek [a,b,*], Dairazalia Sanchez-Cortes [a], Pradeep Rangappa [a], Joshua Hughes [h], Jakub Tkaczuk [a,j], Alejandra Sanchez Lara [a], Driss Khalil [a], Johan Rohdin [b], Dawei Zhu [c], Aravind Krishnan [c], Dietrich Klakow [c], Zahra Ahmadi [d], Marek Kováč [e], Dominik Bobošš [e], Costas Kalogiros [f], Andreas Alexopoulos [f], Denis Marraud [g]

[a] *Idiap Research Institute, Martigny, Switzerland*
[b] *Faculty of Information Technology, Brno University of Technology, Czech Republic*
[c] *Saarland Informatics Campus, Saarland University, Germany*
[d] *L3S Research Center, Leibniz Hannover University, Germany*
[e] *Phonexia, Brno, Czech Republic*
[f] *AEGIS IT Research, Athens, Greece*
[g] *Airbus Defence and Space - Intelligence, Elancourt, France*
[h] *Trilateral Research, Liverpool, United Kingdom*
[i] *University of Zurich (UZH), Zurich, Switzerland*
[j] *Present address: ETH Zurich, Zurich, Switzerland*

ARTICLE INFO

Keywords:
Criminal investigations
Speaker identification
Automatic speech recognition
Network analysis

ABSTRACT

A criminal investigation is a labor-intensive work requiring expert knowledge from several disciplines. Due to a large amount of heterogeneous data available from several modalities (i.e., audio/speech, text, video, non-content data), its processing raises many challenges. It may become impossible for law enforcement agents to deal with large amounts of highly-diverse data, especially for cross-border investigations focused on organized crime. ROXANNE EC H2020 project developed an all-in-one investigation platform for processing such diverse data. The platform mainly focuses on analyzing lawfully intercepted telephone conversations extended by non-content data (e.g., metadata related to the calls, time/spatial positions, and data collected from social media). Several state-of-the-art components are integrated into the pipeline, including speaker identification, automatic speech recognition, and named entity detection. With information extracted from this pipeline, the platform builds multiple knowledge graphs that capture phone and speaker criminal network interactions, including the central network and their clans. After hands-on sessions, law enforcement agents found the Autocrime platform easy to understand and highlighted its innovative, multi-technology functionalities that streamline forensic investigations, reducing manual effort. The AI-powered platform marks a significant first step toward creating an open investigative tool that combines advanced speech, text, and video processing algorithms with criminal network analysis, aimed at mitigating organized crime.

1. Introduction

Organized crime is among the most challenging types of crime to investigate, and it carries a substantial threat to modern society, national and international security. Significant financial flows within organized crime networks provide criminals with access to resources and modern technologies that enable the effective planning, executing and concealing of their criminal activities and extensions of their cross-national networks. In 2009 transnational organized crime generated profit as high as $870 billion (or 1.5% of global Gross Domestic Product), according to the Unitated Nation Office on Drugs and Crime (UNODC) (Pietschmann and Walker, 2011). The latest 2024 report, although not disclosing estimated profits, shows growing trends for specific markets (like cocaine and synthetic drugs (United Nations Office on Drugs and Crime, 2024).

Fighting organized crime remains an important but challenging task for law enforcement agencies (LEAs) in all countries. To analyze crimi-

* Corresponding author.
E-mail address: petr.motlicek@idiap.ch (P. Motlicek).

https://doi.org/10.1016/j.fsidi.2025.301937
Received 29 November 2024; Received in revised form 7 May 2025; Accepted 10 May 2025
Available online 3 June 2025

nal networks, police practitioners must manually collect, research, and connect information from different sources, including but not limited to eyewitnesses, electronic audio (lawfully intercepted conversations), and social media (group communications in closed sessions). During their meticulous and time-consuming work, investigators screen out information from these sources, identify relevant entities and link them through relations and actions into meaningful networks. As reported by the LEAs involved in the ROXANNE project, between 50 to 80% of the workload per investigative case is dedicated solely to extract preliminary information from raw data such as audio files obtained from telephone wiretaps and other intercepted electronic communication. The situation gets more challenging with the complexity and size of cross-border criminal cases with multimodal data. The larger the case, the more data has to be analyzed. The relations between telephone numbers and individuals become less certain due to the random use of prepaid cell phones, public phones, or shared phones. Multilingualism also adds to the complexity on top of all the above. In these situations, the workload often exceeds the capabilities of the team assigned to the case.

With technologically capable and savvy criminals in mind, the increase of speed and efficiency of investigation through its automation became the main goal of this work. Automating the investigation process can greatly impact the daily work of police practitioners and the resources needed to monitor criminal organizations. Due to the heterogeneous, multi-modal and often multilingual nature of the data, automating this task is also quite complex. The lack of representation of such realistic data in the training of Artificial Intelligence (AI) models matching the domain further adds to the complexity of the problem.

The use of cloud-based services is not feasible due to strict data regulations. Law enforcement and government agencies impose strict and demanding procedures when employing any technology. These include verifying compliance with data protection or the CLOUD Act - ensuring clarity on where data is stored and processed - validating encryption policies, and establishing agreements with service providers that align with national regulations and legal frameworks (Malik et al., 2024; European Parliament and Council of the European Union, 2016; U.S. Congress, 2018). In contrast, offline forensic tools allow organizations to maintain complete control over their data. Additionally, it eliminates legal complexities by keeping data within a controlled environment, minimizing risks related to jurisdictional conflicts or government access requests.

Several commercially available offline tools offer methods for automated criminal network investigation - they are briefly discussed and compared in the Section 2. These tools help the investigators to identify the key players, paths and relations within the criminal networks. They also allow for determining effective network destabilization and disruption strategies, increasing the effectiveness of the investigation. Some of them provide efficient ways to keep track of the geolocation information. Considering various modalities of automatic analysis, all audio, video and text data become highly interesting for the investigation processes. All these types of data are often lawfully used to uncover the identities of individuals during their communication through telephone networks or the internet. Natural language processing and social network analysis are also used in criminal network analysis, e.g. through several other projects such as the Real-time Early Detection and Alert system for online terrorist Content (Red-Alert). Nevertheless, existing developed tools lack a modular adaptive AI framework and are mostly closed-source software.

This paper presents an open platform referred to as Autocrime, developed under the EC H2020 ROXANNE project[1] aiming to make the analysis of criminal networks faster and more effective. It is developed for police practitioners to process large and heterogeneous data, to automatically extract evidence to assist them in making informative decisions. Autocrime is capable of transcribing audio for the following languages: English, German, Dutch, Greek, Lithuanian, Arabic, and Spanish. The presented, integrated approach combines the findings from different sources to investigate how various processes can mutually support and accelerate each other.

2. Motivation

IoT forensic applications and services in the market focus on data extraction from diverse devices and systems (Android, iOS, WP, etc.). They offer cloud-based, in-premise or hybrid processing services and provide the extracted data in the form of databases, logs or plain files (Almubairik et al., 2025; Mahmood et al., 2024; Oxygen Forensics, 2025; Cellebrite, 2024; MSAB, 2025; Exterro, 2024). They aim to reconstruct activities, event timelines, and gather crucial evidence in criminal investigations. Hardware-based evidence case, is a more challenging task since the access to the functional actual devices is critical (mobile phone, laptop, hard-disk, etc).

After collecting digital evidence, the next challenge is decoding and decrypting the recovered data while identifying its type (Fukami et al., 2024). This includes determining what additional digital evidence may be needed to fill gaps left by unrecovered or encrypted data. In some cases, this may involve requesting data stored on servers and managed by a service provider, subject to legal authorization, so called non-content data (Umberg and Warden, 2014; Karagiannis and Vergidis, 2021; Guild et al., 2021; Avgerinos et al., 2024; Rangappa, P. et al., 2025). Informed decision-making then depends on LEAs ability to integrate multimedia evidence seamlessly while managing the exponential growth of multilingual voice and text-based data.

We present Autocrime as an analytics platform for criminal investigation of telephone networks. Forensic investigations involve large amounts of data spanning different modalities: telephone calls as audio, their metadata, text messages, images and videos from confiscated mobile devices, etc. Telephone calls obtained from Communication Service Providers (CSP) are associated with a significant amount of metadata including the locations from which calls were made, phone numbers, IMEI of devices, etc. The data can be in one of many inter-compatible, electronic, tabular formats such as an Excel file, or a Comma-separated Value (CSV) file. Rapid, automatic processing of such data severely eases an investigator's burden, as manual introspection of all data is prohibitively time-consuming.

To our knowledge and based on the feedback from LEAs participating at various stages of the project, current-day tools are typically unimodal. For instance, the i2 Analyst Notebook tool (N. Harris Computer Corporation, 1990) helps to visualize telephone networks, and its commercial speaker recognition software only compare voices across audio files. However, such tools are not designed to suit the common workflow of crime investigators, where the data needs to be processed and visualized in a logical order, further continue the analysis with new data which arrive to be processed, allowing for error-correction of outputs from technologies built on machine learning approaches (which is typically the case for text, audio and video processing). The motivation behind our platform arises from the lack of modern tools that allow effective visualization and interaction with multiple modalities, and also serve as rather a tool in the loop of the conventional investigative procedure.

Table 1 compares the features of the Autocrime platform against other fee-based tools (unimodal and multimodal) used by LEAs. Watson Natural Language Understanding (NLU) (IBM, 2024) is a text based tool for entities, keywords, sentiment, emotion, relations and syntax; which supports 13 languages. Google Cloud - Natural Language AI (Google) is a structured and unstructured text-based tool that allows sentiment analysis, entity recognition (people, places and events). It also supports Optical Character Recognition (OCR) for scanned documents and video and speech-to-text in more than 135 languages. ACU-EXPERT (Acustek, 2009) is an audio analysis solution that allows recording device identification, speaker voice identification, speaker diarization, speech en-

[1] https://www.roxanne-euproject.org.

Table 1
Autocrime platform features versus other investigation tools used by the Law Enforcement Agencies.

Features	Autocrime	Other tools
General		
Combines two (or more) cases	Yes	Usually yes
Incorporates several techniques	Yes	Usually yes
Speech processing		
Voice detection	Yes	Usually no
Speaker Diarization	Yes	Usually no
Speaker clustering and identification	Yes	Usually no
Gender identification	Yes	Usually no
Language identification	Yes	Usually no
Automatic Speech Recognition	Yes	Usually no
Keyword-spotting	Yes	Usually no
Natural Language Processing (NLP)		
Named Entity Recognition	Yes	Usually yes
Topic Detection	Yes	Usually no
Identification of unknown second parties	Yes	Usually no
Detection of mentions of third parties	Yes	Usually no
Visual analysis		
Facial Similarity Search	Yes	Usually no
Scene Similarity Search	Yes	Usually no
Network Analysis		
Link Prediction	Yes	No
Social Influence Analysis	Yes	Usually no
Community Detection	Yes	Usually no
Outlier Detection	Yes	No
Cross Network Analysis	Yes	No

hancement and text decoding. MSAB XRY (MSAB, 2025) integrates mobile and other digital data, call detail records, timeline view and geolocation data integration. Adobe Audition (Adobe Systems Incorporated, 1982) is a paid audio toolkit that allows audio processing of various audio file formats, provides noise reduction and speech enhancement. VoiceGain (Voicegain, 2019) offers speech-to-text and natural language processing (NLP), topic detection and sentiment analysis (positive and negative segments). Speechmatics (Speechmatics, 2006) offers speech-to-text, summarization, and translation with support for 50 languages. Newton Technologies (Newton Technologies, 2008) offers to convert audio and video to text, as well as speaker diarization with support for more than 30 languages. The i2 Analyst's Notebook offers processing text files and structured data, people and organizations' social network analysis, and sequence of events; nevertheless it requires connection to other i2 iBase tools (N. Harris Computer Corporation, 1990) for additional features. Tovek (Tovek, 1993) is a platform with forensic tools for structured and unstructured data that provides entity recognition (people, organizations), time analysis and discovery of events and relationships. It also offers speech-to-text and image processing (object recognition). Hasken (2010) is able to process structured and unstructured data (i.e., files, chat logs, browser histories, emails, location, etc.) with metadata from images and events (timelines and maps).

In addition to the input from their regular investigative tools, a survey of the law enforcement agencies from 40 different countries was conducted to identify the most critical problems present in cross-border investigations (ROXANNE Project Consortium, 2022). Full details of the survey compiled in a deliverable are available for the EU civil security practitioners through the European Commission only (Center for Security Studies (KEMEA), 2022). The survey results pointed to a critical issue related to the pricing and scalability of current solutions with increase in data size and number of suspects. Another problem identified by LEAs was that individuals often used multiple phone numbers to avoid identification. Yet another problem is the complexity in analyzing the content of these calls - criminals using abnormal terms to describe their activities or interests. Finally, there is the problem of using closed-source software - not all LEAs could use solutions without accessing and checking the source code first.

Table 2
Data Input to Autocrime required from a LEA.

Audio files	
mp3, wav, sph, flac, etc.	(any file format)
stereo, mono	(any recording format)
Metadata CSV file header	
∗ Audio filename	∗ Intercepted number id
∗ Caller phone number	∗ Receiver phone number
∗ Caller name	∗ Receiver name
∗ Call timestamp	

3. Platform overview

In real data scenarios, mapping speakers from mono-channel recordings to files listing multiple telephone numbers or speaker IDs is significantly challenging due to the lack of automatic mapping. With this motivation, in Autocrime, we manually align speaker data based on audio characteristics and timestamps, converting various WAV formats to standardized mono output. This approach facilitates accurate speaker diarization and ensures reliable association between audio and metadata, thereby improving the analysis and scalability of investigations.

In this section, we describe the platform from the user's perspective.

3.1. User input

Running the platform requires minimal configuration from the user. The input consists of audios and a metadata file, described in Table 2.

3.2. Technology workflow

The data flow of the system is shown in Fig. 1. Once the input is ingested, various speech analysis technologies are run to process and analyze the data: (i) the telephone recordings are converted to 8 kHz wav format; and (ii) Voice Activity Detection (VAD) is applied to locate speech in every audio recording. In case of single-channel (mono),[2] (iii) speaker diarization is executed to identify regions belonging to different speakers.

Eventually, speaker models, also referred to as voice embeddings, are extracted for each audio and subsequently used in speaker recognition (and clustering) engine.

To automatically transcribe telephone conversations, Automatic Speech Recognition (ASR)[3] is executed. Currently, the platform supports 7 languages (see Section 4). Automatic transcriptions generated by ASR are fed to the text mining module (i.e., Named Entity Recognition (NER)) (see Section 4.3.1). NER is currently supported for German and English. Eventually, a mention network is run which enables to identify geographical locations, person mentions, and organization names available in the transcripts (and used by network analysis module).

As part of network analysis, a communication network (i.e., telephone network generated by considering telephone numbers only, and speaker network representing individuals appearing in phone conversations) is constructed. In case of speaker network, the voice embeddings extracted from telephone calls are compared and clustered considering the distance metrics between those embeddings.[4] Additional network analysis algorithms are applied to the communication network, including Community detection (to group speakers based on frequent interactions) and Social influence analysis (to identify the most influential

[2] In case of 2-channel audio, the speakers (caller and callee) are separated into two separated audio tracks).

[3] Often referred as speech-to-text.

[4] Similar voice embeddings are bound to the same individuals composing a single node in the speaker network. Directed network edges are constructed, representing caller and calee.

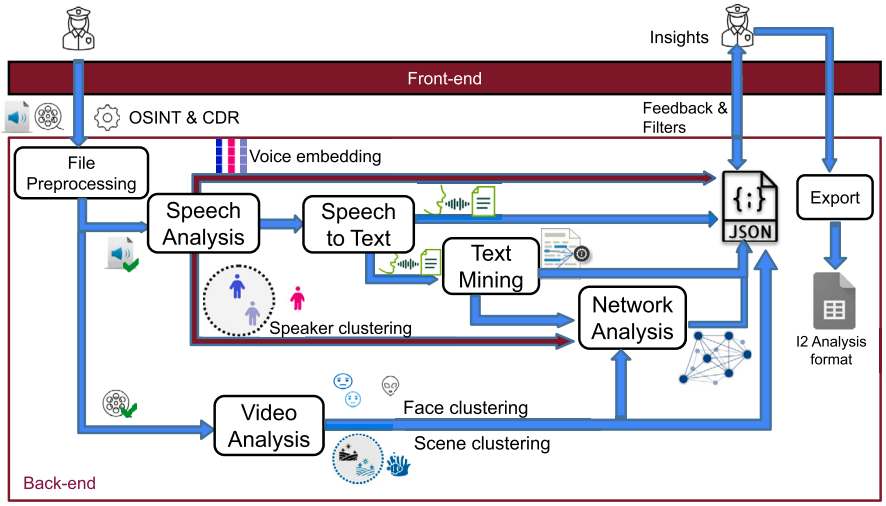

Fig. 1. Schematic representation of the Autocrime analysis workflow.

individuals within the network), see Section 4.5. In addition, scenes from videos (see Section 4.4) can be enrolled onto existing nodes in the speaker-based network or in a new media type node, which will be added to the network automatically.

The results of the Autocrime analysis workflow (i.e., back-end) are available for the human operator using visualisation tool (i.e., front-end) (Network Visualization for Criminal Behavior Analysis for AUTOCRIME and TRACY, 2025). Besides graphically presenting results, the visualisation tool permits to upload new data to the platform on an already processed case. It also allows searching through data and filtering the results. Last but not least, it displays the projected speaker and phone networks and their statistics.

4. Core technologies: a modular approach

The advantage of the platform lies in its modularity: speech, text, image processing, and network analysis technologies are integrated to fit seamlessly within the workflow and data processing chain.

4.1. Data ingestion and preprocessing

In the first step of speech analysis workflow, VAD distinguishes speech segments from non-speech segments in an audio signal. The VAD outputs allow for only the parts of a recording with enough speech signal to be processed by subsequent speech technologies, e.g., speaker recognition (Sec. 4.2.1) or speech recognition (Sec. 4.2.5). In this step, the audio format is also standardized for speech analysis.

4.2. Speech analysis

4.2.1. Speaker recognition (SR)

With standardized audio format, a Speaker Recognition system identifies who is speaking in a given audio recording. This process is an integral part of Autocrime because the identities of speakers in recordings from criminal investigations are usually unknown. As opposed to the controlled laboratory environment, in which datasets to train machine learning models are collected, the nature of data available for investigation is uncontrolled, presenting challenges.

The speaker recognition system follows the state-of-the art approach where the speech from a single speaker is converted by an artificial neural network into a fixed size vector representation, referred to as a voice embedding or speaker embedding (Variani et al., 2014; Snyder et al., 2017). The voice embedding extractor integrated in the platform is based on ResNet (He et al., 2015; Zeinali et al., 2019), an architecture

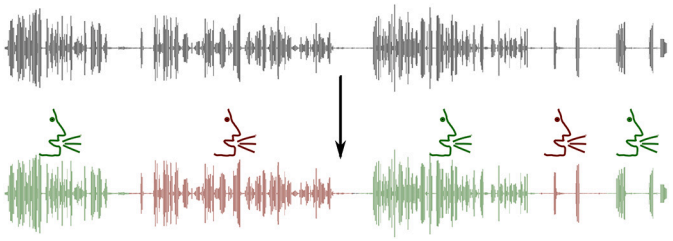

Fig. 2. Schematic representation of speaker diarization inputs (waveform) and results (clusters).

available as part of the VBx recipe (Landini et al., 2020). The embeddings are then modeled by a generative model called Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe, 2006a).

4.2.2. Diarization of mono-channel audio

When the recording has only one channel (i.e., mono recording where both telephone tracks are stored in one channel), diarization is applied to segment the audio into individual speakers as shown in Fig. 2. Diarization is also applied on audio extracted from videos. Variational Bayes Hidden Markov Models (VBx) is used in Autocrime (Diez et al., 2020). The Agglomerative Hierarchical Clustering (AHC) algorithm that is internally used in VBx is applied on speaker embeddings projected in a speaker-discriminative space induced by a pretrained PLDA model (Ioffe, 2006b).

4.2.3. Clustering of unknown speakers

Given a dataset for criminal investigation, each individual data may be: known, unknown but relevant, or unknown but irrelevant to the case. The unknown cases naturally occur when investigators collect communication data of known speakers. An unknown speaker can also appear due to the known speaker using a different, but unverified, telephone number or device for communication.

The platform uses AHC of speaker embeddings obtained from the speaker recognition system. Prior information, for instance two different phone numbers definitely belonging to two different persons, can be added to the clustering algorithm by setting the relevant entries in the similarity score matrix to a very high value enforcing the recordings to be in the same cluster and to a very low value for the recordings to be in different clusters. The complete process is illustrated in Fig. 3.

The investigators are explicitly allowed to merge speakers despite system clustering results, demonstrating forensic-standard practice for error correction (Andersen et al., 2025). Manual corrections help address systematic errors caused by noisy recordings or overlapping speak-

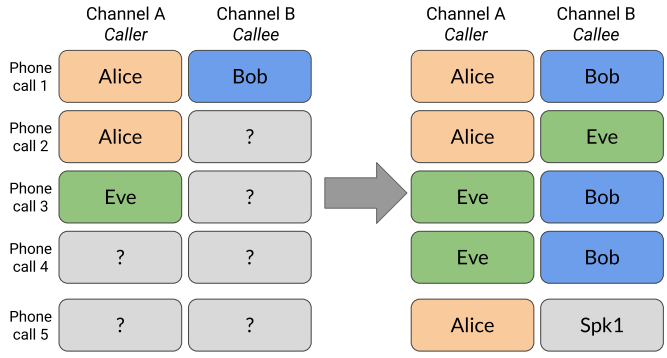

Fig. 3. Schematic representation of speaker clustering with enrollment. Left-hand side shows recordings before clustering - two of which are identified as Alice, one as Eve, and one as Bob. The speakers in the rest of the recordings are unknown. After clustering, Alice is identified in three, Eve in two, and Bob in three recordings. Additionally, one recording is identified as an unknown speaker: "Spk 1".

ers. This aligns with findings from literature, where ASR systems inherently produce errors (e.g., insertions/deletions) that require human validation for reliability (Patman and Chodroff, 2024).

4.2.4. Gender identification

A pretrained Convolutional Neural Network (CNN)-based model trained on the REPERE challenge corpus was used for gender identification from speech and tested on ROXSD (Doukhan et al., 2018). The model was trained on 121 hours of audio data from 2274 speakers. We observed that the CNN-based model performed significantly better than a simple Gaussian mixture model (GMM) in terms of gender classification accuracy (75% with GMMs vs 95% with CNN-based models).

4.2.5. Automatic speech recognition (ASR)

The audio after VAD is transcribed by ASR. Seven languages are supported: English, German, Dutch, Greek, Lithuanian, Arabic, and Spanish. We fine-tuned the wav2vec 2.0 based XLSR-53 model for each of the seven languages (Conneau et al., 2020). For each language, a dedicated vocabulary, Acoustic Model (AM) and Language Model (LM) is used. The transcription is then linked to the speaker, an output of speaker recognition and speaker clustering.

4.2.6. Boosting informative words

Conversational speech collected in criminal networks can have highly specialized vocabulary (e.g. names, nicknames, codewords). Moreover, this can vary significantly across cases. Word boosting provides a simple way to add this prior knowledge during transcription. We use the technology that has already been applied in different domains, such as the air traffic control systems (Kocour et al., 2021). This technology adds value to the system when the investigators have prior knowledge about specific words to pay attention to when generating automatic transcription. This user-driven boosting technique increases the probability of words in the language model, increasing the chance that these words are correctly recognized.

4.3. Text analysis

NLP allows analyzing the output text of the speech technologies, described in Sec. 4.2. Detecting named entities in automatic transcripts from conversational speech is often challenging due to several factors: the lack of textual cues (ASR transcripts often lack punctuation and capitalization, which are typically used by NER models trained on written text (Nguyen and Yu, 2021)); ASR errors (transcription errors in ASR output can propagate to the NER system); informal language (spoken communication is less formal than the typical text data used to train NER systems); context dependency (the meaning of words in conversational speech heavily relies on surrounding context); overlapping voices; and Out-Of-Vocabulary (OOV) words. Autocrime attempts to tackle the above challenges through its modularity and through the use of NLP technologies described below.

4.3.1. Named-entity recognition

The NER module aims at automatically extracting useful named entities from transcripts of the audio files. In Autocrime, NER is performed on the transcripts from the calls obtained with the speech technologies, described in Sec. 4.1. This step assists LEA agents in quickly focusing on the informative pieces of text from vast amounts of textual data and allows them to enhance other components like network analysis. The Autocrime's NER module is built on RoBERTa (Liu et al., 2019) and supports three types of recognized entities: person, location, and time. These three entities enable better contextualization of events - such as determining when they occurred (past, future, or exact timing), where they took place, and who was involved. The LM model is trained with more focus on high-level semantics and less focus on grammar and orthography, i.e., spelling, casing, and punctuation. Duplicate named entities are eliminated by creating a union of all identified entities per transcribed call.

4.3.2. Mention network and co-reference resolution

Telephone conversations contain critical information about places and organizations in an investigation. The Mention Network module identifies people mentioned in the calls, who are not necessarily the speakers in the audio, using the transcripts from the speech recognition module. For example, in the transcribed segment shown in Fig. 4 between Mark (caller) and Paul (callee), the line "I hear John is sick..." appears. Since neither the caller nor the callee is named John, John is identified as a mentioned named entity in the conversation. This further complements the Speaker Network with more edges between parties to indicate name or location mentioned entities.

While the NER module extracts the person-information from a telephone conversation, the co-reference resolution links all linguistic expressions, like pronouns and referrals in a span of text, to the original entities they refer to. These person-mentions can either be the persons in the call or a third party, outside the call. The "mention disambiguation" is splitting the person-mentions into *party* and *third party* entities. It is done by automatically checking the pronouns/mentions in the text and extracting the closest entity token, the positions of other entities, and the surrounding words from the context.

More specifically, Autocrime integrates a hybrid model for the mention disambiguation module. A combination of the rule-based and the co-reference resolution models is used. This hybrid approach emphasizes the rule-based model towards the initial portion of the conversation and moves onto the co-reference model afterward, improving the overall recognition accuracy. As co-reference model, we used NeuralCoref[5] which predicts the co-reference scores for mentions/pronouns and named entities, and later a simple ranking algorithm is used to find the best matching entity.

The results of this process are schematically represented in Fig. 4. Mention disambiguation module is evaluated using Accuracy of the Mention Labeling (AML). AML measures ability of the model to separate mentions i.e., the percentage of detected entities that are correctly identified as *parties* or *third parties*. Any person-entity missed out by the NER and ASR modules cannot be retrieved and processed by the disambiguation module. Therefore, these entities are not considered when computing the AML.

[5] https://github.com/huggingface/neuralcoref.

Fig. 4. Example output of the co-reference resolution model, where all personal pronouns are linked back to their original entities.

4.4. Image and video processing

Autocrime supports processing of image and video data. The platform is equipped with technologies enabling automatic processing of images and videos to enrich speaker networks with additional edges and nodes to support investigations, as shown in Fig. 5. To fulfill the investigators' needs, face identification and scene characterization technologies are integrated in the data processing chain. An advantage of these technologies in Autocrime is their integration into the graphical user interface, where investigators can directly visualize faces within the speaker network. This capability, combined with the ability to access all related visual documents through the node details view, enhances the focus of attention by reducing the need to shift through irrelevant videos.

4.4.1. Face characterization

An existing open-source state-of-the-art face detection and embedding extraction model[6] was utilized. This model is a Pytorch implementation of Arface, which was trained by Deng et al. (2019) and demonstrates 98% accuracy on the benchmark MegaFace dataset (Kemelmacher-Shlizerman et al., 2016). To protect individuals' identities, all images or videos containing faces from the ROXSD dataset were pseudo-anonymized using the face-swapping technologies. We adapted the SimSwap framework[7] on a subset of selfie videos from Realtime Selfie Video Stabilization dataset,[8] utilizing fake faces generated by the StyleGAN2 model[9] to replace the original faces with the non-existent ones. Finally, face matching is only performed between a limited set of manually enrolled face pictures corresponding to primary suspects or victims. These enrolled faces are compared against all ingested images and videos in order to allow limited usage of facial technology in a proportionate manner, aligning with ethical recommendations. The efficacy of the pseudonymization process was assessed by evaluating cosine similarities between original and swapped faces. Results demonstrated successful use of pseudonym names and consistency in face swapping across videos.

To optimize matching performances, we developed an additional face clustering and summarization algorithm to summarize each input video in a set of identities: each identity is described by their K most relevant face observations. This algorithm utilizes the Density Based Spatial Clustering for Applications with Noise (DBSCAN) clustering method (Ester et al., 1996) to group observations of the same face, even across varying head poses. DBSCAN iteratively gathers sample points in clusters through an incremental search for nearest neighbors.

4.4.2. Scene characterization

The scene characterization technology and the matching capability involves describing an image with a signature or embedding that encodes distinctive visual features to differentiate one scene from another. The method utilized the place embedding training based on a ResNet backbone and an ArcFace module initially developed for face recognition but adapted successfully for scene embedding extraction (Deng et al., 2019). The module was trained on the Google Landmark Dataset (Weyand et al., 2020), resulting in a model that achieved the state-of-the-art performance on various datasets, such as the Revisited Paris dataset (Radenović et al., 2018).

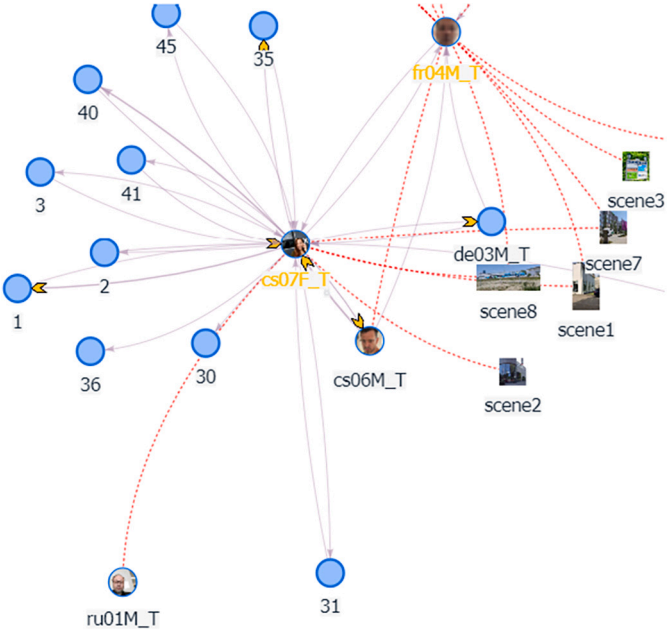

Fig. 5. ROXSD example: image edges added into the speaker network after the ingestion of images and videos. The connection between cs06M_T and fr04M_T nodes is a result of both faces being detected in a common ingested image. In addition, new edges were automatically created between speaker nodes and the enrolled scenes. All referenced edges are displayed in red dotted lines (for interpretation of the colors in the figure(s), the reader is referred to the web version of this article).

The use of scenes in Autocrime is similar to the use of faces: scenes from videos can be enrolled onto an existing speaker node (indicating a relationship between the location and a person) or into a new media type of node, which will be added to the network. Enrolled images are automatically compared against a database of ingested images and videos. Matches are identified based on cosine similarity thresholds defined in a configuration file. This functionality offers investigators immediate access to relevant visual documents without the need to sift through a large volume of material. In addition, these media nodes can be used to facilitate automatic links between speakers and locations.

4.5. Network analysis

Network analysis uncovers hidden patterns in the behavior and relations among individuals in networks. This technology leverages the information extracted by the previous modules (i.e., speech, text, and video). Network analysis technologies presented in (Hagberg et al., 2008) are applied on telephone and speaker networks to:

1. Quantify the influence of individuals within a social network and identify the most influential ones with Social Influence Analysis technology.
2. Identify frequent cohesive subgroups with Community Detection.

[6] https://github.com/foamliu/InsightFace-PyTorch.
[7] https://github.com/neuralchen/SimSwap.
[8] https://github.com/jiy173/selfievideostabilization.
[9] https://github.com/NVlabs/stylegan2.

3. Predict missing, hidden, and unobserved interactions with Link Prediction.
4. Identify individuals who exhibit abnormal interactions with others in the group with Outlier Detection.
5. Find graph nodes representing the same entities across two different networks.

4.5.1. Social influence analysis

This task aims to quantify individuals' global influence over others within a social network. Each individual is assigned a relative importance score to measure their influence on a network. We use Pagerank as the centrality measure. Pagerank score (Brinkmeier, 2006) measures the significance of a target individual by counting the weighted vote it receives from other individuals. Each in-link in the network is considered a "vote", and the vote frequency of other influential individuals determines the weight of the link. Mathematically, this problem is a recursive equation that can be solved using power iteration methods.

As part of the Autocrime project, the social influence analysis method is evaluated using the following approach: each node in the graph (representing an individual) is assigned a ground truth label based on the scenario designed and implemented during data collection (Motlicek et al., 2024). The method then assigns a social influence score to each node. The overall accuracy is computed as a weighted score that reflects the alignment between each node's influence score and its corresponding ground truth label.

4.5.2. Community detection

The individuals within a network tend to form communities, a cohesive group of individuals whose interactions inside the community are more frequent than their interactions with the rest of the network. This community structure is often hidden and not well-defined. Community detection aims to uncover the hidden structures based on the interactions among individuals and associated information, such as the attributes and the metadata.

Community detection applied to criminal networks is relatively unexplored. There are challenges in applying the community analysis technologies to subgroups such as gangs, cliques, families, and clans. Global analysis of gangs often fails to capture the fluidity and contradictions in gang identification across diverse cultural contexts (Fraser and Hagedorn, 2018). Subgroup analysis in social sciences also faces methodological challenges, including issues with sample selection, statistical power, and the risk of false positives due to multiple hypothesis testing (Breck and Wakar, 2021).

Networks may comprise multiple clans for remarkably different reasons: some clans are the offspring of others, other clans are clashing and conflicting; the community analysis aims at detecting the intensified interaction between opposing groups (Calderoni et al., 2017).

We employ the K-clique-based method (Palla et al., 2005) for community detection. The K-clique method discovers closely connected groups within a network. It is known to perform significantly good on real networks scenarios (i.e., sparse graphs); the computation of the algorithm is also fast (Abboud et al., 2024) compared to other methods due to the fairly limited number of cliques (Naik et al., 2022; Gupta et al., 2022). It is also characterized by a robust capability to discriminate between random graphs, reducing the risk of creating artificial subgroups.

The implementation of the methods is individual-centric and parametric (it takes K, often ranging from 2 to 5, as an input). The method finds the overlapping communities and assumes that each individual can be a member of more than one community. Each community is identified by unifying adjacent K-cliques in the network. A K-clique is a fully connected subnetwork, and two K-cliques are adjacent if they have $K-1$ individuals in common.

4.5.3. Link prediction

Criminal investigation procedure tends to reliably retain data on the primary suspect, while details regarding marginal suspects may be filtered out and lost (Berlusconi et al., 2016). Thus, the link prediction methods assists investigators by providing possible new leads in the investigation. Link prediction aims to uncover missing or hidden interactions and relations in the network; and to predict the most probable interactions to be formed in the near future.

We use node centrality-based prediction (Ahmad et al., 2020) to find the nodes with higher centrality degree that may have a higher chance of forming new links. More specifically, given an individual u (target) in the network, we would like to identify the top k other individuals who are not directly connected with u but have a very high probability of having interacted.

4.5.4. Outliers detection

In network data, outliers might be individuals who exhibit abnormal interactions and/or relations with others. Given an input network, we rank individuals by anomalous behavior and present only the top-ranked individuals. Outlier detection techniques have found extensive application in various domains such as fraud detection (Bolton and Hand, 2002), crime investigation (Lin and Brown, 2006) and more recently in identifying terrorist activities (Wang and Li, 2019).

Specifically for the Autocrime platform, outlier detection is conducted by analyzing and thresholding individuals' social influence scores (previously described in Section 4.5.1). For evaluation purposes, the person of interest in the ROXSD criminal case (i.e., the suspect) is labeled as the "Target" individual. In contrast, other individuals appearing in the case (e.g., food order call responders, uninvolved family members, or paid confidential informants assisting the police) are classified as "NonTarget" outliers.

4.5.5. Cross-network entity matching

The cross-network entity matching, also called "Cross-Domain Entity Resolution" or "Entity Linkage", finds graph (i.e., network) nodes that indicate the same entities (i.e., individuals) across different graphs. The goal of the technology is to perform node matching between two networks. These two networks can be results from different criminal cases or can be obtained as two networks considering two complementary graphs. In the latter case, the first network is represented by ROXANNE speaker network (i.e., network of individuals established from telephone conversations), while the second network can represent social media, where the user-names are often anonymous.

In Autocrime, cross-network entity matching is performed using Term Frequency-Inverse Document Frequency (TF-IDF), an algorithm originally developed for NLP. TF-IDF is used to assess the importance of a word within a document relative to a collection of documents. In context of Autocrime, each "word" represents an attribute of a graph node (such as a speaker's name from phone conversations, a nickname or username from social media, or other entity names mentioned in conversations). The "document" corpus corresponds to a graph representing a specific modality, such as telephone calls or social media posts. The TF-IDF algorithm processes JSON data from two networks and compares it against a ground truth (CSV) file that specifies known node matches.

5. Operational use

To demonstrate Autocrime's functionality, we present a scenario from a LEA agent X, using the Autocrime platform.

LEA X is investigating a drug dealing criminal network in Europe, depicted as a multinational organization. Members are suspected to speak their native languages and English. To follow this investigation, 432 phone calls are lawfully intercepted by national police organizations, including call data records (i.e., telephone numbers, date, and time) from the service provider, alongside the audio recordings. As expected, the

calls contain a conversation between two (or more) suspects, i.e., intergroup communication and commercial activity. Among all the calls, there are some made between two suspects, a suspect and a non-suspect, and between two non-suspect persons. Furthermore, calls are known to have background noise, change of speakers and language in the middle of a conversation. Gathered evidence from some suspects resulted in collecting several additional messages, including relevant images and videos.

In traditional day-to-day work, LEA X had: (1) to invest a lot of time (more than 7 hours) identifying calls manually and categorizing it as informative or non-informative (e.g., an answering machine); (2) to request a budget between 432 - 4,212 € (prices 1-3.25 € per minute for the case of German) to delegate the manual transcription of the audios to a service provider, knowing that the cost can increase depending on the turnaround time required (adding proofreading, editing, and quality checks), estimated in more than 33 hours. (3) to pay 15.55 € in the best case scenario (0.036 € per minute) using cloud services, if validated and GDPR compliant for its use (where the data is going to be stored, processed, etc), validated data encryption policy, and signed agreements with the service provider respecting regulations and laws of the country (Malik et al., 2024).

Instead, LEA X could use Autocrime to save time, money, and to focus only on the following relevant questions to obtain insights from the case:

How to automatically identify informative calls, speakers in the audio and its content? Autocrime requires a configuration file with the call data logs and the folder containing the calls. It then automatically identifies speakers in a conversation (diarization), identifies the language spoken (language identification) and transcribes (ASR) the content in the recognized language. If the same speaker is identified in several calls (speaker recognition), Autocrime labels this consistently along with the information from the call logs, e.g., the respective telephone number associated.

Who is the caller/the receiver? Are there recurrent names or nicknames? Autocrime can differentiate between individuals (nodes) by assigning them one of the call parties (caller or receiver), or a third party (person entity) who is mentioned (name or nickname) in the call (performed by Mention Disambiguation)

What is the topic of the call (without having to listen to it entirely? Autocrime currently classifies a phone conversation with one of the following labels: Drugs, Work Conversations, Family-Friend Conversations, Money, Meeting, Other.

Who calls whom, when, and from which number? Who is in the center? Are there clans? After processing call conversations, Autocrime builds two graphs: the phone-based graph and speaker-based graph. The phone-based graphs focus on phone numbers as provided from the call records. The speaker-based graph focuses instead on individuals involved in the conversations with their respective assigned identifiers (or names if available). Both graphs display most central and cohesive subgroups (Community Detection), identifies individuals who exhibit abnormal interactions (Outlier Detection) and displays the details of the calls (date, time, etc.).

Does the conversation reveal who, when or where transactions are carried out? Although the full content of a conversation is available, LEA X can focus on relevant information revealing when or where. Geopolitical entities are highlighted (countries, cities, states) as well as non-geopolitical entities such as hotels, streets, and restaurant names, among others (location entity). Also, absolute dates and times are highlighted (time entity) as shown in Fig. 6 .

Is the same person in the images and videos, as the one speaking on a drug dealing phone conversation? Autocrime can perform face recognition through images and videos. Audio from videos is extracted and voices of speakers are compared with previous voices processed in the case. Similarly, environments are identified in both images and videos. The resulting matches are displayed in the graphs linked to the identified speakers and associated numbers as shown in Fig. 5 .

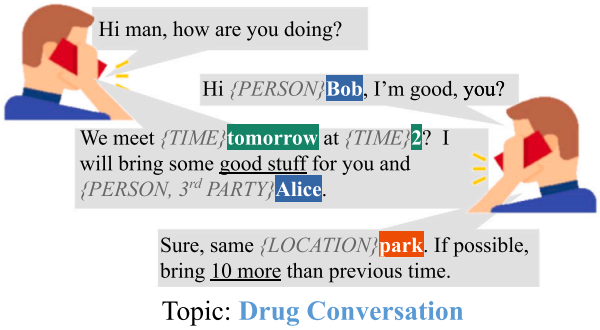

Fig. 6. A call conversation processed by Autocrime. Entities PERSON, LOCATION and TIME are highlighted, a third-party PERSON (Mention Disambiguation) is also identified in the conversation and the topic Drug Conversation is inferred.

Towards the end of the ROXANNE project (October 2022), a hands-on session was organized to allow LEAs, researchers, policy makers, representatives from industry and civil society organizations, to experiment with the Autocrime functionalities. In total 59 in-person and 67 remote attendees participated (including 11-Roxanne LEA partners). After the session, some participants expressed interest in integrating more technologies (like geoposition) into the platform and expanding its language processing capabilities. Other participants highlighted the platform's uniqueness in integrating multiple technologies from various modalities while simultaneously speeding up LEAs investigations that would otherwise require manual effort.

6. Evaluation of technologies

6.1. Dataset: ROXSD

The ROXANNE project's main objective was to support police officers, investigating large criminal cases, by offering conventional machine learning technologies processing multimodal data (speech, text, video) and later on post-processed using network analysis. However, evaluation challenges arose due to scarce and restricted real-world criminal activity datasets. To address them, the ROXSD dataset (Motlicek et al., 2024) was introduced. ROXSD is a set of wiretapped telephone conversations (collected in a usual way through a communication service provider), simulating a scenario of cross-border drug trafficking prepared by LEAs. ROXSD, freely available for security practitioners in Europe,[10] comprises ~20 hours of telephone and video conversations involving 104 speakers, with ground-truth annotations for precise evaluation and technology development. Emphasizing multimodality and multilinguality, the dataset includes metadata and prior knowledge like suspects' biometric profiles. ROXSD serves as a pivotal resource for advancing technology in criminal research, particularly in speech, text, and network analysis domains. It fills the gap left by the absence of comprehensive real-world datasets, showcasing the potential of simulated datasets in organized crime analytics.

6.2. Evaluation metrics

One of the key factors of Autocrime is its ability of measuring the quality of implemented technologies in a reproducible manner. This section therefore presents the evaluation metrics used to objectively assess the Autocrime technologies on publicly available ROXSD data. The metrics are divided in four categories: Speech, NLP, Network Analysis, and Visual Analytics.

[10] https://www.roxanne-euproject.org/data.

Table 3
Summary of telephone conversations collected as part of ROXSD data used for evaluating speech components of the Autocrime platform.

Category	Details
# Total telephone calls	464
# Stereo calls	461
# Mono calls	3
# Calls used for diarization evals	364[1]
# Calls used for ASR evals	481
Total duration of calls	18 h 28 min

[1] The dataset includes 3 mono calls (i.e., the calls are recorded as 1-channel recording), and 361 stereo calls (i.e., each stereo recording is converted to a mono recording) for the evaluation of speaker diarization.

6.2.1. Speech

Following speech technologies were evaluated:

- VAD: The *Detection Error Rate* (der-VAD) is used as a metric to evaluate the performance of VAD. When applied at the frame level, it measures the rate of misclassification of voice activity. Alternatively, it can be assessed at the segment level, where the accuracy is determined based on whether the center of each segment is correctly classified.
- Speaker diarization: the performance is measured using a standard *Diarization Error Rate* (DER) metric. DER is calculated as the combined total of three types of errors: False Alarms (FA) (incorrectly detecting speech where there is none), missed speech detections (failing to detect actual speech), and speaker label confusion (incorrectly attributing speech to the wrong speaker).
- Speaker and gender identification: the accuracy metrics is used, i.e., a measure of how well the system correctly identifies speakers or gender from a set of possibilities.
- ASR: Automatic Speech Recognition (ASR) is assessed using Word Error Rate (WER) for both English and German languages to objectively assess the difference between the transcribed ASR output and the reference (human obtained) transcript. Evaluating ASR systems across different languages is a complex task due to the need for accounting for diverse linguistic features, accents, and contextual variations. Due to these challenges, Autocrime's ASR performance is evaluated on English and German datasets, as expanding the evaluation to other languages would have required addressing these additional complexities. The ASR technology is designed to be multilingual and is capable of supporting the evaluation of additional languages, such as: Greek, Spanish, Arabic, Dutch, and Lithuanian. Nevertheless, these languages were not included in the evaluation phase. The primary reason for their exclusion was the lack of available datasets that are sufficiently comprehensive and standardized to support a rigorous evaluation.
- Speaker clustering: to evaluate a speaker clustering system, one needs a mapping between the speaker labels assigned by the system and the ground truth labels. This mapping can be obtained with the Hungarian algorithm Kuhn (1955) to maximize the clustering accuracy, defined as a percentage of correctly recognized recordings. Mapping is only used when evaluating the system against a ground truth reference and not in real operation, where the ground truth is unknown.

Table 3 provides a summary of key information from the ROXSD telephone conversations used to evaluate the speech components of Autocrime. It includes information on the number of telephone recordings, stereo/mono format, availability of the groundtruth reference (i.e., speaker label), as well as the total duration of the set for ASR evaluation.

6.2.2. NLP

To assess the performance of the NER, standard *F1-score* (i.e., for classification tasks, it is the harmonic mean of precision and recall) is used. In case of Mention Network (i.e., the relationship between a pair of entities), we use *Accuracy* quantifying whether detected entities were correctly identified as *parties* or *third parties*. More specifically, the task of NER is inherently complex, especially when dealing with a multilingual dataset and diverse types of entities. The NER model, deployed in Autocrime, was designed to identify three distinct categories of entities: person names (PERSON), locations (LOCATION), and time-related expressions (TIME) across both English and German texts. The challenge lies not only in correctly identifying and categorizing these entities within a single language, but also in adapting the model to maintain consistent performance across two different linguistic structures. This complexity is compounded by the varied contexts in which these entities appear. Person names, for example, can differ significantly in structure and frequency between languages, while locations might be referenced in culturally specific ways that require the model to have a nuanced understanding of both languages. Time expressions, which can vary significantly in format and contextual meaning between English and German, add another layer of difficulty.

The ROXSD dataset poses an added challenge due to the presence of a large number of unique entities to be recognized by NER system. The English subset contains 1,293 entities, of which 417 are unique, while the German subset includes 527 entities with 230 unique entities. This diversity in the dataset increases the task's complexity, as the model must be able to identify and classify a wide array of entity types correctly, despite potential ambiguities and inconsistencies in the data.

In Autocrime, the integration of NLP technologies is naturally applied to the automatic transcriptions of telephone calls generated by ASR systems. NER and Mention Network techniques are evaluated across three scenarios:

1. Groundtruth transcriptions: human-annotated transcriptions of telephone conversations.
2. ASR transcriptions: automatically generated transcriptions by ASR systems, which may contain errors.
3. Boosted ASR transcriptions: ASR outputs enhanced using a predefined list of named entities, incorporated through lattice rescoring Kocour et al. (2021) to improve ASR (and subsequently NER) accuracy.

6.2.3. Network analysis

Six network analysis technologies were integrated into the Autocrime platform, supporting two types of networks: telephone and speaker networks.

Telephone network is based on Call-Detail Records (CDRs) (i.e., metadata file required to execute Autocrime technologies described in Section 3.1). The network presents all wiretapped telephone numbers; the ones originating, receiving calls to/from and the interactions represented as edges. Telephone network allows to analyze social influence (i.e., measuring the importance of each entity based on the topology of the social network). The complexity of the telephone network is highlighted by its size: it contains 115 nodes (representing unique phone numbers) and 236 edges. Analyzing such a network involves navigating a moderately dense graph, where identifying significant patterns requires accounting for the interactions among a substantial number of entities.

The speaker network is derived from the Autocrime outputs (specifically from automatically clustering speakers in ROXSD telephone conversations and from detecting mentions in automatically generated ASR transcripts by NER). Two types of nodes are defined (connected with directional edges): circular nodes (individuals identified in raw telephone recordings by speaker clustering) and triangular nodes (corresponding to "names" mentioned during conversations). The directional edges connect pairs of nodes that indicate direct telephone calls between

Table 4
Evaluation of speech technologies.

Technology	Metric	Performance (%)
VAD	der-VAD	1.2
Speaker Diarization	DER	14.8
Speaker clustering	Accuracy	92.0
Closed set Speaker ID	Accuracy	95.0
Open set Speaker ID	Accuracy	93.0
Gender ID	Accuracy	95.0
ASR English	WER	28.4
+ boosting	WER	28.5
ASR German	WER	35.9
+ boosting	WER	35.9

Table 5
Performance results for boosting of 55 unigrams in ROXSD English subset. The boosting algorithm is applied during ASR inference by using lattice rescoring.

	Baseline	Lattice re-scoring
# true positives	598	640
# false positives	98	108
# false negatives	599	557
F1-Score (PERSON)	15.4%	20.6%
F1-Score (LOCATION)	44.0%	47.9%
F1-Score (TIME)	78.4%	78.8%
Average F1-Score	45.9%	49.1%

individuals (purple solid lines) and represent mentions of names during the phone calls (blue solid lines). The speaker network is notably larger, with 127 nodes (representing unique speakers) and 578 edges. The increased size and complexity of this network requires advanced techniques to manage and interpret the dense connections and interactions within it.

Several network analysis algorithms are developed and integrated in Autocrime:

- (1) Community detection to identify cohesive groups of individuals (whose intra-group interaction is denser and more frequent than their interaction with the rest of the network).
- (2) Social influence analysis, which is applied on both phone and speaker network.
- (3) Link prediction, to predict the existence of a link between two entities in a network.
- (4) Outlier detection, to identify entities that differ from the rest ones, which could indicate either a person of high or low interest.
- (5) Cross-network matching to improve identification of persons of interest which might appear in different networks (e.g., in speaker network created from telephone conversations and in auxiliary network built from social media connections).

For all technologies, we use either *F1-score* (i.e., correctly or incorrectly detected community) or *Accuracy*. In case of the link prediction, top-5 accuracy is measured, assuming that the true label only needs to appear in the top 5 predictions (the 5 classes with the highest predicted probabilities).

6.2.4. Visual analytics

Visual analytics technologies, including Face Detection and Scene Characterization, were evaluated using *Precision* and *Recall*, i.e., the proportion of correctly detected faces or scenes while taking into account the number of false alarms (in case of precision) and the number of misses (in case of recall).

6.3. Results

This section details the performance of the technologies achieved by Autocrime on ROXSD data. The results are summarized in corresponding Tables 4 – 8, and are discussed below in terms of their implications and effectiveness.

6.3.1. Speech technologies

The results of VAD performance, evaulated on the ROXSD data are shown in Table 4. The der-VAD reaches 1.2%, compared to 4.1% of the WebRTC engine.[11]

The diarization framework integrated in Autocrime operates at a DER of 14.8%.

Speaker clustering with the standard AHC algorithm requires a score threshold as an argument to determine when to stop the clustering process. By adjusting this threshold, the user can control the number of clusters (i.e., unique speakers) produced. Using a specific threshold, 88 clusters were obtained, corresponding to 96 unique speakers, with a clustering accuracy of 92%.

The closed-set Speaker Identification (SID) achieved an accuracy of 95%, indicating high reliability when the speakers are within a known set. Open-set scenario refers to a situation where the SID system must identify whether a given speech sample belongs to one of the known speakers (speaker already enrolled) or if it comes from an unknown speaker (a speaker the system has not encountered before). The open-set SID was simulated with 13 pre-enrolled speakers. It is expected that the accuracy decreases, in this case, from closed-set SID (95%) to 93%.

For gender identification, the task achieved a high accuracy of 95.0%, as shown in Table 4.

Autocrime integrates a universal multilingual ASR model (i.e., encoder pre-trained on large amounts of unlabeled data), further trained on target monolingual sets for seven languages.[12] In Table 4, we show performance (in WER) for English (28.5%) and German (35.9%) on ROXSD data. Although WERs are relatively high (especially due to many specific but informative entities appearing in the data), the ASR transcripts are available for further processing. The performance for these informative name entities are further improved by boosting.

ASR boosting of apriori known named entities during inference time, can increase the probability of correct recognition, as presented in Table 5. The presented results are obtained by directly analyzing the ASR transcripts, without deploying the NLP module on top of them. More specifically, the boosting technology used in Autocrime is lattice re-scoring. The lattices are created during decoding and taken as input. The scores of acoustic and language models are split since the process only rescores the language modeling part and takes the acoustic scores as is. The re-scoring step extracts the n-best list from lattices and updates the sentence scores of the n-best list with new ones, corresponding to increased weight for the informative monograms available in the lexicon. As compared with the baseline, improvement is significant (up to 5%) for person based recognition from 15.4% to 20.6%.

6.3.2. NLP

As shown in Table 6, NER system achieves the F1-score of 82.8% and 70.1% for English and German ROXSD subsets, respectively. The F1-scores drop significantly when using automatically (ASR) generated transcripts (i.e., 39.7% and 27.9% for English and German entities, respectively. Nevertheless, word (uni-gram) boosting deployed during ASR inference can improve NER accuracy by 3-4% absolute on both languages. The Mention Network approach achieves the accuracy of 74.8% when applied on manually annotated speech transcripts. In case of using automatically generated ASR transcripts, the accuracy slightly decreased, while it was further increased with word boosting.

[11] https://webrtc.org.

[12] For mode details see the README: https://github.com/idiap/autocrime.

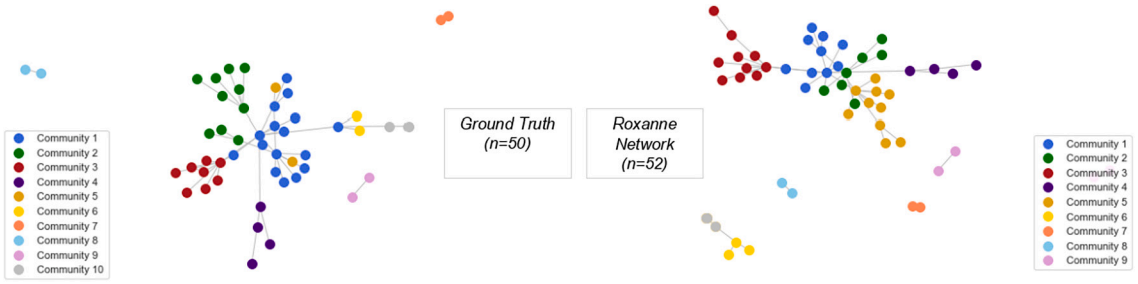

Fig. 7. Evaluation set of the language communities of individuals in the ROXSD network (ground truth on the left and the Autocrime results on the right).

Table 6
Evaluation of NLP technologies.

Technology	Metric	Input source	Performance (%)
NER English	F1-Score	Text	82.8
		ASR	39.7
		+ boosting	43.2
NER German	F1-Score	Text	70.1
		ASR	27.9
		+ boosting	30.6
Mention network	Accuracy	Text	74.8
		ASR	71.6
		+ boosting	75.3

Table 7
Evaluation of network analysis technologies.

Technology	Metric	%
Community Detection	F1-Score	32.8
Phone-Social Influence	Accuracy	79.5
Speaker-Social Influence	Accuracy	95.0
Link Prediction	Top-5 Accuracy	58.8
Outlier detection	Accuracy	100.0
Cross-network matching	Top-1 Accuracy	75.0

Table 8
Evaluation of visual analytics technologies.

Technology	Metric	Performance (%)
Face detection	Recall	98
	Precision	100
Scene & Similarity match	Recall	70
	Precision	86

6.3.3. Network analysis

The F1-score of 32.8% is obtained for community detection, as shown in Table 7, suggesting that while it could identify some community structures, the overall precision and completeness were limited, given the ROXSD scenario. Specifically as pointed by Fig. 7, community detection primarily groups individuals based on the language they use during the call. This is logical for the international criminal casework, where individuals from the same community (e.g., drug trafficking in Prague) often use several languages (as it was simulated through ROXSD scenario).

An accuracy of 79.5% and 95.0% is achieved for social influence for telephone and speaker network, respectively, indicating strong performance in speaker network. The Top-5 Accuracy for link prediction is 58.8%, reflecting a moderate ability to predict the most likely connections within the network. The most prominent result is the Outlier Detection; it achieved a perfect accuracy of 100.0%, demonstrating full effectiveness in identifying anomalies within the network, at least when evaluated on the ROXSD data. Identified phone/speakers outliers reflect indeed infrequent interactions within their community.

Eventually, the Top-1 Accuracy results indicate a solid performance in correctly matching nodes across different networks for cross-network node matching with a 75.0% of Accuracy.

6.3.4. Visual analytics

Table 8 shows the results for detected faces and scene characterization. The face detection algorithm demonstrated excellent performance with a recall of 98% and a precision of 100%, ensuring that nearly all relevant faces were detected with no false positives.

Scene characterization and similarity matching achieved a recall rate of 70% and a precision of 86%, demonstrating strong performance in identifying and matching similar scenes.

7. Privacy-by-design

The concept of Privacy-by-Design places privacy concerns at the center of technology development. As a concept, it is routed in privacy enhancing, or privacy-respecting technologies (Chaum, 1985).

Privacy-by-Design was developed by the former Information and Privacy Commissioner of Ontario, Dr. Ann Cavoukin who suggested 7 foundational principles:

1. Proactive not reactive; preventative not remedial. Anticipate, identify and prevent privacy invasive events before they occur. Privacy issues need to be considered early in the design process, including the stage where initial ideas and objectives are being considered (Wright and Hert, 2012).
2. Privacy as the default setting. Build in the maximum degree of privacy into the default settings for any system or business practice.
3. Privacy embedded into design. Embed privacy settings into the design and architecture of information technology systems and business practices instead of implementing them after the fact as an add-on.
4. Full functionality - positive-sum, not zero-sum. Accommodate all legitimate interests and objectives in a positive-sum manner to create a balance between privacy and security.
5. End-to-end security - full lifecycle protection. Embed strong security measures to the complete lifecycle of data to ensure secure management of the information from beginning to end.
6. Visibility and transparency - keep it open. Assure stakeholders that privacy standards are open, transparent and subject to independent verification.
7. Respect for user privacy - keep it user-centric. Protect the interests of users by offering strong privacy defaults, appropriate notice, and empowering user-friendly options.

Taking these principles into account during the technology design process enables the privacy of end-users and people who might be affected by the technology to be considered and respected as much as can reasonably be achieve during design. Doing so can alleviate concerns of people who would normally be concerned about a loss of privacy and engender trust from those people into the product that is being developed.

Privacy-by-Design does have its limitations, particularly in the context of technologies developed for LEA investigations, which by their nature may infringe on the privacy of suspects. During the design process, technology developers should take steps to build tools that support privacy protections and enable compliance with data protection regulations. When these technologies are used, legal obligations to protect individuals' personal data fall on end-users, who act as data controllers.

In the case of Autocrime, the project partners considered data protection legislation during the design phase to ensure compliance throughout development and to support compliance by end-users. Nevertheless, Privacy-by-Design does not guarantee full compliance, and LEAs will still need to implement appropriate measures to ensure the protection of personal data during the use of Autocrime.

8. Platform Availability and License

Autocrime is freely available for any active LEA, academic partners recognized to be active in security-related researchers and R&D in civil security from EU. The license type used in this project is Apache License 2.0, version January 2004. Interested parties can visit the main website of the project for more details.

Further results from evaluation of the Autocrime platform using ROXSD internal data as well as other (publicly available) data to assess individual AI technologies (for speech, text, video and network analyses) can be found at the GitHub repository.

9. Summary

This paper discusses the development of various speech, NLP, and video technologies within the ROXANNE project. It builds upon previous work in these areas, offering methods for LEAs to enhance their current technological capabilities. The focus is on leveraging information from different sources to understand the structure of criminal networks.

The primary aim of the report is to present a streamlined pipeline that exploits multi-modal information, in particular, from phone calls but also from videos. We first extract metadata such as phone numbers and geo-location which can provide an initial social network structure assuming there is a one-to-one mapping between person and phone numbers. The audio processing starts with voice activity detection that separates speech from non speech in the audio. If needed, diarization may then be used to separate multiple speakers in the recording. Speaker recognition can then be used to produce a more accurate network than that based on phone numbers. Depending on circumstances, the phone number information can also be used to improve the speaker recognition performance. Speech recognition is then used to produce a transcript of the speech. In this process it is possible to boost the prediction of important words specified by the user. We then build a "mention network" from the transcripts using entities mentioned in conversations. Thereafter, we process the transcripts to find third parties mentioned in the network and integrate it into the network. Thus, we investigate social, textual and acoustic information from a network to provide LEAs with meta-information about the crime.

We furthermore present the work done in integrating video technologies into this pipeline. An exploratory analysis is also made into using geo-location information in the paradigm proposed. This should provide the reader with insights into the potential of previously untapped information resources and display the extent to which the retrieval and analysis of this information can be automated.

The discussed evaluation results show a strong foundation. The methodology could be further expanded across diverse linguistic and cultural settings, incorporating LEAs feedback to refine adaptability and ensure relevance in different cultural contexts.

As part of future work, we will fine-tune the pre-trained CNN model (Doukhan et al., 2018) for gender identification using a specialized dataset of conversational speech data to further improve its performance on ROXSD. Once optimized, we will integrate the model into the Autocrime platform, enabling more accurate and efficient processing of speech data within the platform's framework.

Other enhancements include the addition of latest NER models (currently highly performing even at Zero-shot (Zhu et al., 2025; Rafique et al., 2024)) for their corresponding ASR engines, including: Dutch, Greek, Lithuanian, Arabic, and Spanish. Offering as well, detection of a larger variety of entities including nationalities or religious/political groups, companies, institutions, objects, vehicles, etc.

Planned upgrades identified by the consortium members includes geolocation, based on availability of geolocation data (from mobile devices, call logs from a service provider, videos or other wearable devices). The visualization of maps indicating points of interest or where/when detected events happened, would provide LEAs with a wider source of evidence in their investigations.

Clearly, improvements and addition of components will require validation of several tests to ensure a seamless functionality of all the components. More importantly, it is critical to have a discussion with LEAs, to determine languages for each country based on local user needs, prioritizing common and relevant languages while avoiding unnecessary additions.

Autocrime brings streamlined, efficient, and agile framework, enabling flexible component integration and replacement for a high-performing and robust service, without the overhead of a monolithic back-end.

CRediT authorship contribution statement

Srikanth Madikeri: Writing – original draft, Methodology, Software, Validation, Investigation. **Petr Motlicek:** Writing – review & editing, Project administration, Supervision, Funding acquisition, Conceptualization. **Dairazalia Sanchez-Cortes:** Writing – review & editing, Investigation, Validation. **Pradeep Rangappa:** Methodology, Software, Validation, Writing – review & editing. **Joshua Hughes:** Investigation, Data curation, Ethics. **Jakub Tkaczuk:** Methodology, Software, Validation, Writing – review & editing. **Alejandra Sanchez Lara:** Visualization, Software, Data curation. **Driss Khalil:** Methodology, Software, Validation, Writing – review & editing. **Johan Rohdin:** Methodology, Software, Validation, Writing – review & editing. **Dawei Zhu:** Methodology, Software, Validation, Formal analysis NLP technologies. **Aravind Krishnan:** Methodology, Software, Validation, Formal analysis NLP technologies. **Dietrich Klakow:** Methodology, Software, Validation, Formal analysis NLP technologies. **Zahra Ahmadi:** Methodology, Software, Validation, Formal analysis NLP technologies. **Marek Kováč:** Investigation, Data curation, Ethics. **Dominik Boboš:** Methodology, Software, Validation, Writing – review & editing. **Costas Kalogiros:** Software, Visualization, Writing – review & editing. **Andreas Alexopoulos:** Software, Visualization, Writing – review & editing. **Denis Marraud:** Methodology, Software, Validation, Formal analysis NLP technologies.

Declaration of competing interest

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

Acknowledgement

The research described in this work was performed within the framework of the Project ROXANNE – Real time netwOrk, teXt and speaker ANalytics for combating orgaNized crimE. This project received funding from the European Union's Horizon 2020 Work Program for research and innovation 2018-2020, under grant agreement No. 833635. The work was also partially supported by TRACY project (Big-data analytics from base-stations registrations and cdrs e-evidence system) funded under Digital Europe Programme (DIGITAL) under grant agreement No. 101102641.

ROXSD Data availability

ROXSD data (Motlicek et al., 2024) collected as part of the ROXANNE project and used for objective evaluations of the Autocrime platform is freely available for civil security researchers from the EU at https://www.roxanne-euproject.org/data.

References

Abboud, A., Fischer, N., Shechter, Y., 2024. Faster combinatorial k-clique algorithms. In: Latin American Symposium on Theoretical Informatics. Springer, pp. 193–206.

Acustek, 2009. Forensic page. Acustek - technical surveillance, tscm and forensic solutions. https://acustek.com/forensic/forensic-audio/. (Accessed 27 January 2024).

Adobe Systems Incorporated, 1982. Adobe audition. https://www.adobe.com/products/audition.html. (Accessed 29 January 2024).

Ahmad, I., Akhtar, M.U., Noor, S., Shahnaz, A., 2020. Missing link prediction using common neighbor and centrality based parameterized algorithm. Sci. Rep. 10, 364.

Almubairik, N.A., Khan, F.A., Mohammad, R.M., Alshahrani, M., 2025. Wristsense framework: exploring the forensic potential of wrist-wear devices through case studies. Forensic Sci. Int. Digit. Investig. 52, 301862. https://doi.org/10.1016/j.fsidi.2025.301862. https://www.sciencedirect.com/science/article/pii/S2666281725000010.

Andersen, D.B., Sunde, N., Porter, K., 2025. Tool induced biases? Misleading data presentation as a biasing source in digital forensic analysis. Forensic Sci. Int. Digit. Investig. 52, 301881. https://doi.org/10.1016/j.fsidi.2025.301881. https://www.sciencedirect.com/science/article/pii/S2666281725000204.

Avgerinos, N., Mertis, P., Tsagaris, M., Desipris, N., Lyberopoulos, G., Theodoropoulou, E., Filis, K., Sarajlić, J., Pastor, M., Chatzakou, D., et al., 2024. Innovative digital forensic and investigation tools for law enforcement: the empower & tracy approach. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, pp. 80–93.

Berlusconi, G., Calderoni, F., Parolini, N., Verani, M., Piccardi, C., 2016. Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis. PLoS ONE 11, e0154244. https://doi.org/10.1371/journal.pone.0154244. https://dx.plos.org/10.1371/journal.pone.0154244.

Bolton, R.J., Hand, D.J., 2002. Statistical fraud detection: a review. Stat. Sci. 17, 235–255.

Breck, A., Wakar, B., 2021. Methods, challenges, and best practices for conducting subgroup analysis. OPRE Rep 17. https://acf.gov/opre/report/methods-challenges-and-best-practices-conducting-subgroup-analysis.

Brinkmeier, M., 2006. Pagerank revisited. ACM Trans. Internet Technol. 6, 282–301. https://doi.org/10.1145/1151087.1151090.

Calderoni, F., Brunetto, D., Piccardi, C., 2017. Communities in criminal networks: a case study. Soc. Netw. 48, 116–125. https://doi.org/10.1016/j.socnet.2016.08.003. https://www.sciencedirect.com/science/article/pii/S0378873316300363.

Cellebrite, 2024. Why cellebrite - cellebrite. https://cellebrite.com/en/why-cellebrite/.

Center for Security Studies (KEMEA), 2022. Deliverable d2.3 end-user requirements. In: Confidential Deliverable of the EU Horizon Project ROXANNE, available upon request through the European Commission only, https://cordis.europa.eu/project/id/833635/results.

Chaum, D., 1985. Security without identification: transaction systems to make big brother obsolete. Commun. ACM 28, 1030–1044. https://doi.org/10.1145/4372.4373.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M., 2020. Unsupervised cross-lingual representation learning for speech recognition. ArXiv preprint arXiv:2006.13979.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699.

Diez, M., Burget, L., Landini, F., Cernocky, J., 2020. Analysis of speaker diarization based on Bayesian HMM with eigenvoice priors. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 355–368. https://doi.org/10.1109/taslp.2019.2955293.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., Meignier, S., 2018. An open-source speaker gender detection framework for monitoring gender equality. In: Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on. IEEE, pp. 5214–5218.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd, pp. 226–231.

European Parliament and Council of the European Union, 2016. General data protection regulation (gdpr). In: Regulation (EU) 2016/679.

Exterro, 2024. FTK forensics toolkit - digital forensics software tools. https://www.exterro.com/digital-forensics-software/forensic-toolkit.

Fraser, A., Hagedorn, J.M., 2018. Gangs and a global sociological imagination. Theor. Criminol. 22, 42–62. https://doi.org/10.1177/1362480616659129. https://journals.sagepub.com/doi/10.1177/1362480616659129.

Fukami, A., Buurke, R., Geradts, Z., 2024. Exploiting rpmb authentication in a closed source tee implementation. Forensic Sci. Int. Digit. Investig. 48, 301682. https://doi.org/10.1016/j.fsidi.2023.301682.

Google Cloud natural language | Google cloud. https://cloud.google.com/natural-language. (Accessed 27 January 2024).

Guild, E., Kuskonmaz, E., Mitsilegas, V., Vavoula, N., et al., 2021. Data retention and the future of large-scale surveillance: the evolution and contestation of judicial benchmarks. In: Queen Mary Law Research Paper.

Gupta, S.K., Singh, D.P., Choudhary, J., 2022. A review of clique-based overlapping community detection algorithms. Knowl. Inf. Syst. 64, 2023–2058.

Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring network structure, dynamics, and function using networkx. In: Varoquaux, G., Vaught, T., Millman, J. (Eds.), Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, pp. 11–15.

Hasken, 2010. The open digital forensic platform. Home - hansken. https://www.hansken.nl/. (Accessed 29 January 2024).

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. https://arxiv.org/abs/1512.03385. https://doi.org/10.48550/ARXIV.1512.03385.

IBM, 2024. IBM Watson natural language processing. https://www.ibm.com/products/natural-language-processing. (Accessed 30 January 2025).

Ioffe, S., 2006a. Probabilistic linear discriminant analysis. In: Computer Vision – ECCV 2006. Springer Berlin Heidelberg, pp. 531–542.

Ioffe, S., 2006b. Probabilistic linear discriminant analysis. In: European Conference on Computer Vision. Springer, pp. 531–542.

Karagiannis, C., Vergidis, K., 2021. Digital evidence and cloud forensics: contemporary legal challenges and the power of disposal. Information 12, 181.

Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E., 2016. The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873–4882.

Kocour, M., Veselý, K., Blatt, A., Gomez, J.Z., Szöke, I., Černocký, J., Klakow, D., Motlicek, P., 2021. Boosting of contextual information in ASR for air-traffic call-sign recognition. In: Proc. Interspeech 2021, pp. 3301–3305.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. In: Naval Research Logistics Quarterly, pp. 83–97.

Landini, F., Profant, J., Diez, M., Burget, L., 2020. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. https://arxiv.org/abs/2012.14952. https://doi.org/10.48550/ARXIV.2012.14952.

Lin, S., Brown, D.E., 2006. An outlier-based data association method for linking criminal incidents. Decis. Support Syst. 41, 604–615.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: a robustly optimized bert pretraining approach. https://doi.org/10.48550/ARXIV.1907.11692.

Mahmood, H., Arshad, M., Ahmed, I., Fatima, S., ur Rehman, H., 2024. Comparative study of iot forensic frameworks. Forensic Sci. Int. Digit. Investig. 49, 301748. https://doi.org/10.1016/j.fsidi.2024.301748. https://www.sciencedirect.com/science/article/pii/S2666281724000672.

Malik, A.W., Bhatti, D.S., Park, T.J., Ishtiaq, H.U., Ryou, J.C., Kim, K.I., 2024. Cloud digital forensics: beyond tools, techniques, and challenges. Sensors 24. https://doi.org/10.3390/s24020433. https://www.mdpi.com/1424-8220/24/2/433.

Motlicek, P., Dikici, E., Madikeri, S., Rangappa, P., Jánošík, M., Backfried, G., Thomas-Aniola, D., Schürz, M., Rohdin, J., Schwarz, P., Kováč, M., Malý, K., Boboš, D., Leibiger, M., Kalogiros, C., Alexopoulos, A., Kudenko, D., Ahmadi, Z., Nguyen, H.H., Krishnan, A., Zhu, D., Klakow, D., Jofre, M., Calderoni, F., Marraud, D., Koutras, N., Nikolau, N., Aposkiti, C., Douris, P., Gkountas, K., Sergidou, E., Bosma, W., Hughes, J., Team, H.P., 2024. Roxsd: the roxanne multimodal and simulated dataset for advancing criminal investigations. In: The Speaker and Language Recognition Workshop (Odyssey 2024), pp. 17–24.

MSAB, 2025. Mobile forensics solutions for forensic specialists - MSAB. https://www.msab.com/roles/forensic-specialists/.

N. Harris Computer Corporation, 1990. i2 group. Link analysis software: discover, create and exploit actionable intelligence. https://i2group.com/. (Accessed 27 January 2024).

Naik, D., Ramesh, D., Gandomi, A.H., Babu Gorojanam, N., 2022. Parallel and distributed paradigms for community detection in social networks: a methodological review. Expert Syst. Appl. 187, 115956. https://doi.org/10.1016/j.eswa.2021.115956. https://www.sciencedirect.com/science/article/pii/S0957417421013099.

Newton Technologies, 2008. Newton technologies. https://www.newtontech.net/en/. (Accessed 29 January 2024).

Nguyen, M., Yu, Z., 2021. Improving named entity recognition in spoken dialog systems by context and speech pattern modeling. In: Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Singapore and Online, pp. 45–55. https://aclanthology.org/2021.sigdial-1.6.

Oxygen Forensics, 2025. Oxygen forensic® detective - all-in-one solution. https://www.oxygenforensics.com/en/products/oxygen-forensic-detective/.

Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818. https://doi.org/10.1038/nature03607.

Patman, C., Chodroff, E., 2024. Speech recognition in adverse conditions by humans and machines. JASA Express Lett. 4, 115204. https://doi.org/10.1121/10.0032473. https://pubs.aip.org/jel/article/4/11/115204/3320022/Speech-recognition-in-adverse-conditions-by-humans.

Pietschmann, T., Walker, J., 2011. Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes: research report, Technical Report, United Nations Office on Drugs and Crime. http://www.unodc.org/documents/data-and-analysis/Studies/Illicit_financial_flows_2011_web.pdf.

Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O., 2018. Revisiting Oxford and Paris: large-scale image retrieval benchmarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5706–5715.

Rafique, K.A., Pansuriya, M., Wawrzik, F., Grimm, C., 2024. Decoding domain-specific ner: a performance evaluation of chatgpt, bi-lstm, and bert. In: 2024 11th International Conference on Machine Intelligence Theory and Applications (MiTA), pp. 1–8.

Rangappa, P., Muscat, A., Lara, A.S., Motlicek, P., Antonopoulou, M., Fourfouris, I., Skarlatos, A., Avgerinos, N., Tsangaris, M., Kostka, K., 2025. Detecting criminal networks via non-content communication data analysis techniques from the Tracy project. Digital Forensics and Cyber Crime 613. https://doi.org/10.1007/978-3-031-89363-6_20.

ROXANNE Project Consortium, 2022. Introduction to the autocrime platform. In: Presented at the ROXANNE Final Conference. 29 November 2022, Paris, France. https://www.roxanne-euproject.org/results/files/roxanne-final-conference_introduction-to-autocrime-platform.pdf, 2022.

Sanchez Lara, A., Motlicek, P., Sanchez-Cortes, D., Rangappa, P., Madikeri, S., Khalil, D., 2025. Idiap Research-Report 2025, Technical Report, Idiap-RR, Martigny. https://publications.idiap.ch/publications/show/5598.

Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S., 2017. Deep neural network embeddings for text-independent speaker verification. In: Interspeech 2017. ISCA, pp. 999–1003. http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0620.html.

Speechmatics, 2006. Ai speech technology | speech-to-text api | speechmatics | home. Speechmatics. https://www.speechmatics.com. (Accessed 29 January 2024).

Tovek, 1993. Data?! We know how .. tovek. https://www.tovek.cz/. (Accessed 29 January 2024).

Umberg, T., Warden, C., 2014. The 2013 Salzburg workshop on cyber investigations: digital evidence and investigatory protocols. Digit. Evid. Elec. Signat. L. Rev. 11, 128.

United Nations Office on Drugs and Crime, 2024. World Drug Report 2024, Technical Report, United Nations, Vienna. https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2024.html.

U.S. Congress, 2018. Clarifying Lawful Overseas Use of Data Act (Cloud Act), Pub. L. No. 115-141, 132 Stat. 348.

Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Florence, Italy, pp. 4052–4056. http://ieeexplore.ieee.org/document/6854363/.

Voicegain, 2019. Speech-to-text apis | speech recognition | voice ai | asr. https://www.voicegain.ai/. (Accessed 29 January 2024).

Wang, P.C., Li, C.T., 2019. Spotting terrorists by learning behavior-aware heterogeneous network embedding. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2097–2100.

Weyand, T., Araujo, A., Cao, B., Sim, J., 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2575–2584.

Wright, D., Hert, P.D. (Eds.), 2012. Privacy Impact Assessment. Springer, Netherlands.

Zeinali, H., Wang, S., Silnova, A., Matějka, P., Plchot, O., 2019. But system description to voxceleb speaker recognition challenge 2019. https://arxiv.org/abs/1910.12592. https://doi.org/10.48550/ARXIV.1910.12592.

Zhu, L., Wang, J., He, Y., 2025. LlmLink: dual LLMs for dynamic entity linking on long narratives with collaborative memorisation and prompt optimisation. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (Eds.), Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics, Abu Dhabi, UAE, pp. 11334–11347. https://aclanthology.org/2025.coling-main.751/.