ROXSD: The ROXANNE Multimodal and Simulated Dataset for Advancing Criminal Investigations

Petr Motlicek1,3, Erinc Dikici2, Srikanth Madikerti1, Pradeep Rangappa1, Miroslav Jánošik2, Gerhard Backfried2, Dorothea Thomas-Aniola2, Maximilian Schürtz2, Johan Rohdin3, Petr Schwarz3, Marek Kováč3, Květoslav Malý3, Dominik Boboší3, Mathias Leibiger5, Costas Kalogiros6, Andreas Alexopoulos6, Daniel Kudenko7, Zahra Ahmadi7, Hoang H. Nguyen7, Aravind Krishnan8, Dawei Zhu8, Dietrich Klakow8, Maria Jofre9, Francesco Calderoni9, Denis Marraud10, Nikolaos Koutras11, Nikos Nikolau12, Christiana Apostiti13, Panagiotis Douris13, Konstantinos Gkountas13, Eleni Sergidou14, Wauter Bosma14, Joshua Hughes15, Hellenic Police Team16

1 Idiap Research Institute, Martigny, Switzerland, 2 HENSOLDT Analytics GmbH, Austria, 3 Brno University of Technology, Czech Republic, 4 Phonexia, Czech Republic, 5 ZITiS, Germany, 6 Aegis IT Research, Germany, 7 University of Hannover, Germany, 8 University of Saarland, Germany, 9 Transcrime, Universita Cattolica del Sacro Cuore di Milano, Italy, 10 AIRBUS Defence and Space, France, 11 ADDITESS Cyprus, 12 ITML, Greece, 13 KEMEA, Greece, 14 Netherlands Forensic Institute, 15 Trilateral Research Ltd, England, 16 Hellenic Police Team, Greece

{petr.motlicek, srikanth.madikerti, pradeep.rangappa}@idiap.ch

Abstract

The ROXANNE project, conducted under the European Union’s Horizon 2020 Programme, aimed to revolutionize criminal investigations by integrating speech, language, and video technologies with criminal network analysis. Despite the success in technology development, the project faced evaluation challenges due to the scarcity and legal restrictions surrounding real-world criminal activity datasets. In response, we introduce ROXSD, a simulated dataset of communication in organized crime. ROXSD is a set of wiretapped conversations (collected through communication service providers) between drug dealers operating suspects, following a realistic screenplay (incl. realistic conditions and constraints of a real investigation) prepared by Law Enforcement Agencies (LEAs). With a focus on multimodality and multilinguality, the dataset comprises 20 hours of telephone and video conversations involving 104 speakers, and is further aligned with ground-truth annotations for each modality involved, enabling precise evaluation and development of technologies. In addition, the multimodal data are enhanced with metadata and prior knowledge (e.g., suspects’ biometric profiles) which is typically available as a result of lawfully intercepted communication. This paper introduces ROXSD as a pivotal resource for advancing technology in criminal research (specifically in domain of speech, text and network analysis). ROXSD not only facilitates in the domain of technology development and evaluation but also showcases the potential of simulated datasets in advancing the field of organized crime analytics, emphasizing the importance of such datasets in the absence of comprehensive real-world alternatives.

1. Introduction

ROXANNE (“Real-time network, text, and speaker analytics for combating organized crime”) was a research and innovation project funded by the European Union’s Horizon 2020 Programme, running from 2019 to 2022. Its aim was to enhance criminal investigations through the extraction of information coming from data in multiple modalities (audio, images and video, and additional traffic data) which are typically accessed from communication service providers. A novel platform was developed which combined speech, language and video technologies as well as criminal network analysis in a flexible and modular workflow. The platform is composed of numerous components running state-of-the-art technologies such as speaker diarization, speaker clustering and identification, automatic speech recognition, named entity recognition, mention disambiguation, topic detection, face detection, scene characterization, social influence analysis, outlier detection, community detection and link prediction.

The core technologies of the ROXANNE platform rely heavily on machine learning approaches in which large amounts of data are required to accurately train the associated models. The data-driven models in the ROXANNE platform were mostly trained using standard publicly available data. To objectively assess the platform, however, the available datasets were found insufficient for a number of reasons: (i) the (very small number of) data sets related to criminal activities which are publicly available contain only a fraction of the modalities of interest in ROXANNE. Therefore, demonstrating the interplay of different technologies was not possible; (ii) data collected from open sources (i.e., accessible through the web and social media platforms or collected by other research projects) had various legal and ethical issues which prevented their use; (iii) the project’s LEA partners were prohibited from transferring any lawfully intercepted data (such as wire-tap recordings, images and videos from their criminal cases) to the technical partners due to legal, privacy and operational reasons.

The ROXANNE Simulated Dataset (ROXSD) was created as a consequence of the drawbacks mentioned above. It is a simulated dataset of typical communication in networks of organized crime. It was developed using standard publicly available data and other datasets collected by LEAs.

https://www.foxanne-euproject.org/platform

Simulated in the sense that all recordings were made with actual
organized crime to facilitate technology development, evaluation and demonstration activities in criminal research. The dataset is built upon the principles of multimodality and multilinguality, and special attention was paid to provide the ground-truth annotations of all modalities involved.

This paper is organized as follows: Section 2 provides an overview of previous activities and projects, Section 3 details how ROXSD was designed and created, and then introduces the ROXSD scenario. The following Sections 4-6 are dedicated to the three main components of ROXSD, namely the calls, video and text subsets, respectively. Section 7 touches upon the ethical measures taken during the data collection process. Section 9 concludes this paper with a summary and some remarks.

The ROXSD is freely accessible for European civil security researchers and developers as well as LEAs through the ROXANNE website [4].

2. Relevant Datasets prior to ROXANNE

The number of datasets related to organised crime which are publicly accessible are rather limited. In [2], the authors introduce covert networks, a collection of past cases’ networks that are collected and made publicly available. This includes a subset with networks among the involved individuals available together with some of their attributes (i.e., name, gender and role). Notable examples include the networks of financial flows from the Madoff fraud case[4] and the network of terrorists involved in the 9/11 attack. In a very recent work, [3] present a burglary dataset provided by Israel National Police. This dataset comprises approximately 30,000 solved burglary cases recorded between 2012 and 2022 in Israel. It includes critical data such as anonymized crime and offender identifiers, timestamps, the number of offenders involved in each case, and the crime site’s min-max scaled coordinates, which prevent accurate location retrieval. It also provides embeddings for case summaries, descriptions of stolen items, and victim testimonies if available. These datasets refer to criminal activities of various types; however, their relational information is limited to contacts and other forms of interactions. An early version of the dataset was described in [4]. This paper describes the final dataset having more data and modalities.

3. Data Collection Setup

As introduced in Section 1, the need for collecting a dedicated dataset originated from the necessity to test and evaluate all of the ROXANNE technologies (individually as well as jointly) together with their interplay in the ROXANNE platform, and to be able to demonstrate the results and outcomes in a realistic criminal investigation scenario without having to deal with legal and ethical issues that could affect research activities involving sensitive data. As such, ROXSD was designed and built on three essential pillars:

- **Realism**: The data resemble what is typically collected in a real investigation as much as possible, both qualitatively and contextually. Actual law enforcement equipment (i.e., established through a communication service provider (CSP)) was used to collect the data.
- **Multimodality**: The data contain multiple data types (modalities) including speech, images/video and text.
- **Multilinguality**: The communication between individuals is carried out in multiple languages, in native and accepted forms.

3.1. Planning and Execution

The initial story (screenplay) of ROXSD was based on a draft scenario prepared by the National Drug Headquarters (NPC) of the Police of the Czech Republic. This draft scenario did not exactly match any real cases, but was inspired by several cases and their own professional experience[4]. Based upon this draft, the main characters and roles in the story, their motivation and relationships were defined in more detail by a task force, composed of several partners of the ROXANNE project.

The collection of data was carried out in three phases between 2020-2022. At each phase, the ROXSD scenario was extended with a follow-up story involving more characters, more data modalities, and more challenging conditions in terms of data quality and communication behavior. In its final version, ROXSD contains 104 individuals and data in speech, video, text modalities as well as call data records. We present the details of data collection for each corresponding data modality in Sections 4-6 below.

Voluntary members of the ROXANNE project partners participated in the data collection process. No biometric data other than the participants’ voices were collected during the data recordings and only simulated (fake) personal data were associated to the characters in the ROXSD story. The participants were asked to choose a non-real name for the characters which they role-played, roughly matching their own nationality (hence, their accent). They were given a summary of the conversation and a number of keywords to mention, but otherwise were left free to structure their conversations. Some participants prepared the conversation script in advance while others chose to improvise. All conversations took place during a set of recording sessions, at various (indoor and outdoor) locations. Participants playing the main characters were accompanied by an assistant who helped them with the communication content and performed quality assurance along the way.

Besides creating the scenario and collecting the data, a great amount of effort was spent for the preparation of accompanying metadata information and ground-truth annotations. This includes, for instance, manually transcribing the audio, determining the speaker and language labels for all segments, manually transcribing the conversations, and annotating the named entities and topics. Special attention was paid to make sure that the metadata set is coherent with the fictional story and the improvised content, and that it is also consistent within itself.

3.2. ROXSD Scenario

The ROXSD story is built on a fictional drug dealing setup in which a group of criminals communicate with each other over the telephone by calls, text and video messaging as well as over the web. Their phone calls are intercepted (wire-tapped) by several (fictional) police organizations. The data also include

To get an insight about the range of criminal activities dealt with by NPC, the interested reader is encouraged to read their public annual reports, available in Czech and English: [5].

---

3.3. Realistic Scenario

Based on the feedback from LEA’s, we also created a scenario that matched typical behavior captured by real criminal network. In this scenario, criminals are found to be exchanging sim cards rather than the telephone numbers. We modified the metadata such that only two of the enrolled speakers have two telephone numbers, and all other enrolled speakers have only one telephone number. Their country code is preserved from the previous scenario. Any unknown speakers are assigned a free phone number available or a random number with a fixed country code. If there are multiple speakers in one side of the call, phone numbers of one of the speakers is assigned randomly.

4. Calls Subset

The calls subset of ROXSD contains audio recordings of the telephone calls made within the (fictional) criminal network. The majority of calls contain a conversation between two (or more) criminals, so called a target call. But there are also non-target calls made between a criminal and an innocent (non-target) character, or between two innocent characters. The topic of each call was preset and the calls were realized by voluntary participants who role-played those characters.

In the first phase of the data collection, Twilio Cloud Communication Platform was used to mimic the interception mechanism. The platform establishes the connection between the communicating parties and stores a stereo recording of each call along with the caller/receiver and date/time information. For the second and third phases, the ROXANNE Consortium had the chance to use real interception equipment. In cooperation with a CSP and with legal permission, two mobile phones and an SIP phone were intercepted. The mobile phones were used by two main characters (criminals) in the story whereas the SIP phone was used as a shared device by several people, some of whom were being investigated as suspects. The remaining participants used their own telephone (mobile/landline) devices. The usage of real equipment also allowed obtaining the Call Data Records (CDR) from the service provider alongside the audio recordings. All real identifiers captured as part of the CDR, including the telephone numbers, date/time information, etc., were pseudonymized.

The phone calls were collected with the objective of creating recordings that resemble a real, natural conversation between the communicating parties. The detailed scenario was not disclosed and the topics of the non-target calls were not enforced.

Some of the recordings in the dataset were especially set up to represent calls of an exceptional nature, such as multiple callers from one mobile device, acoustic noise in the background, change of speakers and/or language in the middle of a conversation, etc. During the course of the phone call, any keywords which were spontaneously uttered during the conversation (e.g. references to third parties) were manually marked and later entered into metadata.

4.1. Data Size and Statistics

The ROXSD calls subset contains 432 intercepted telephone conversations recorded into 481 audio files (recordings), encoded in 8kHz, 16-bit, stereo wave format.

The data is composed of different types of calls: standard phone calls in which the caller dials the receiver’s telephone number, teleconference calls in which the caller calls a third person while already talking to the receiver, and calls that are made to a web conferencing service (Zoom, Webex) where the callers dial a common telephone number (the service’s dial-in number) in order to talk to each other.

There are more recordings (481) than number of calls (432) because some calls were intercepted multiple times from different sides of the conversation: 270 calls are intercepted only on the caller’s side, 111 are intercepted only on the receiver’s side, and 86 are intercepted on both sides.

The phone calls were collected with the objective of creating recordings that resemble a real, natural conversation between the communicating parties. The detailed scenario was not disclosed and the topics of the non-target calls were not enforced.

Some of the recordings in the dataset were especially set up to represent calls of an exceptional nature, such as multiple callers from one mobile device, acoustic noise in the background, change of speakers and/or language in the middle of a conversation, etc. During the course of the phone call, any keywords which were spontaneously uttered during the conversation (e.g. references to third parties) were manually marked and later entered into metadata.

4.1. Data Size and Statistics

The ROXSD calls subset contains 432 intercepted telephone conversations recorded into 481 audio files (recordings), encoded in 8kHz, 16-bit, stereo wave format.

The data is composed of different types of calls: standard phone calls in which the caller dials the receiver’s telephone number, teleconference calls in which the caller calls a third person while already talking to the receiver, and calls that are made to a web conferencing service (Zoom, Webex) where the callers dial a common telephone number (the service’s dial-in number) in order to talk to each other.

There are more recordings (481) than number of calls (432) because some calls were intercepted multiple times from different sides of the conversation: 270 calls are intercepted only on the caller’s side, 111 are intercepted only on the receiver’s side, and 86 are intercepted on both sides.

The phone calls were collected with the objective of creating recordings that resemble a real, natural conversation between the communicating parties. The detailed scenario was not disclosed and the topics of the non-target calls were not enforced.

Some of the recordings in the dataset were especially set up to represent calls of an exceptional nature, such as multiple callers from one mobile device, acoustic noise in the background, change of speakers and/or language in the middle of a conversation, etc. During the course of the phone call, any keywords which were spontaneously uttered during the conversation (e.g. references to third parties) were manually marked and later entered into metadata.

4.1. Data Size and Statistics

The ROXSD calls subset contains 432 intercepted telephone conversations recorded into 481 audio files (recordings), encoded in 8kHz, 16-bit, stereo wave format.

The data is composed of different types of calls: standard phone calls in which the caller dials the receiver’s telephone number, teleconference calls in which the caller calls a third person while already talking to the receiver, and calls that are made to a web conferencing service (Zoom, Webex) where the callers dial a common telephone number (the service’s dial-in number) in order to talk to each other.

There are more recordings (481) than number of calls (432) because some calls were intercepted multiple times from different sides of the conversation: 270 calls are intercepted only on the caller’s side, 111 are intercepted only on the receiver’s side, and 86 are intercepted on both sides.

The phone calls were collected with the objective of creating recordings that resemble a real, natural conversation between the communicating parties. The detailed scenario was not disclosed and the topics of the non-target calls were not enforced.

Some of the recordings in the dataset were especially set up to represent calls of an exceptional nature, such as multiple callers from one mobile device, acoustic noise in the background, change of speakers and/or language in the middle of a conversation, etc. During the course of the phone call, any keywords which were spontaneously uttered during the conversation (e.g. references to third parties) were manually marked and later entered into metadata.

4.1. Data Size and Statistics

The ROXSD calls subset contains 432 intercepted telephone conversations recorded into 481 audio files (recordings), encoded in 8kHz, 16-bit, stereo wave format.

The data is composed of different types of calls: standard phone calls in which the caller dials the receiver’s telephone number, teleconference calls in which the caller calls a third person while already talking to the receiver, and calls that are made to a web conferencing service (Zoom, Webex) where the callers dial a common telephone number (the service’s dial-in number) in order to talk to each other.

There are more recordings (481) than number of calls (432) because some calls were intercepted multiple times from different sides of the conversation: 270 calls are intercepted only on the caller’s side, 111 are intercepted only on the receiver’s side, and 86 are intercepted on both sides.

The phone calls were collected with the objective of creating recordings that resemble a real, natural conversation between the communicating parties. The detailed scenario was not disclosed and the topics of the non-target calls were not enforced.

Some of the recordings in the dataset were especially set up to represent calls of an exceptional nature, such as multiple callers from one mobile device, acoustic noise in the background, change of speakers and/or language in the middle of a conversation, etc. During the course of the phone call, any keywords which were spontaneously uttered during the conversation (e.g. references to third parties) were manually marked and later entered into metadata.
and 45 calls on both sides. There is an additional teleconference call which was interrupted 10 times.

This results in some of the recordings being very similar in content. However, they are not an exact copy of each other, because of the following reasons: (i) The interception begins on the caller’s side as soon as the caller finishes dialing the receiver’s telephone number. Hence, the ring tone as well as any sounds/speech which the caller’s phone picks up before the connection is established are captured by the intercepted recording coming from the caller’s side. For the same reason, the caller’s intercepted recording is a few seconds longer than that of the receiver’s. There are also cases where, although both sides are intercepted, the receiver’s phone is not reachable, therefore there is no recording of the receiver. In such cases, either the receiver’s voice box message or the operator’s out-of-reach message can be heard in the caller’s recording. (ii) For teleconference calls involving three parties, a new interception is initiated when the caller calls the third party in order to connect them into the existing conversation. (iii) For web conferencing where multiple parties call the same (operator) telephone number, each party’s interception begins when they join the conference room. (iv) The audibility of speech in both recordings can be different from each other due to the actual phones used, the background or microphone noise introduced by each side, or issues with the equipment intercepting that side. These inexact copies of the same phone conversation are intentionally preserved in the dataset in order to reflect the nature of interception in the real world.

A call is considered to be successful if the intercepted recording contains the full conversation between the caller and the receiver. There are some failed calls in which a connection cannot be established between the parties, mostly containing only operator or a voicemail box message. Finally, a dropped call happens when the call is picked up but the connection or recording is cut off before the parties finish their conversation. The latter is mainly caused by technical problems related to the data collection and interception process. The failed and dropped calls are also kept in the dataset, as such artefacts represent a realistic depiction of the data acquired in a typical investigation.

The total duration of the 481 recordings is 18h 28min. The shortest call is 0 sec long (failed call) and the longest call is 12min 15sec long. With the help of an automatic voice activity detector (VAD), the total duration of speech activity is calculated to be 19h 34min. This difference occurs partly due to the buffer of silence (sometimes close to 1 second) which is left by the VAD system on both ends of a speech segment, and partly due to the fact that the speakers often interrupt each other in a telephone/teleconference conversation, resulting in overlapped speech when both channels are concerned. Figure ?? shows the histogram of recordings with respect to their audio duration.

There are 104 speakers (characters) in the ROXSD calls subset. 25 of them are target persons, 60 are non-target persons and 18 are unknown characters (including operators, etc.). Apart from the teleconference and web conference calls which contain three or more speakers, there are also calls in which the caller or the receiver hands over the phone to another person. In some other calls, the entire conversation is carried out by a speaker who is not the owner of that phone number. Hence, the speakers in a particular call may not be the regular speakers/owners of the telephone numbers making that call.

In addition to these languages there are a few single phrases or brief sentences uttered in Spanish and Croatian. The subset exhibits a multilingual nature not only across the calls, but also within the calls. In several recordings, the conversation starts in one language and then continues in a different language (code switching). In others, the language of conversation changes when the phone is handed over to another person. There are also calls where the speakers switch the language all of a sudden for a brief moment (typically only for a few sentences, mostly to talk about "business"), and then switch back again.

### 4.2. Metadata

The ROXSD calls subset is complemented with an extensive set of ground truth information, which will be referred to as "metadata". The information in the metadata set is pseudonymized wherever possible, and in other cases, manually prepared / determined based on the circumstances of the fictional story. The metadata set can be semantically grouped into four categories: Speaker and Device Metadata, Call Metadata, Transcripts, and NLP Annotations.

The **Speaker and Device Metadata** category consists of speaker-related and device-related attributes of the characters (speakers) in ROXSD. For each speaker, information such as character name and label, gender, telephone number and device type are provided.

The **Call Metadata** category consists of information which could be obtained by the interception system through the telecommunication operator via the Call Data Records (CDR) and the Base Transceiver System (BTS) records. For each audio file, information such as filename, telephone number of the caller/receiver, IMEI of the caller/receiver, date and time of the call, languages spoken in the call are provided.

The **Transcripts** category contains manual transcriptions of the conversation for each call (audio recording) in their original language. In order to obtain the transcripts, first the stereo audio recordings were split into single (mono) channels. Each channel was sent through a VAD to mark the begin-end timestamps of the boundary of the segments containing speech. Speaker and language labels were added to each detected segment, in such a way that a single segment is composed of a single speaker speaking a single language. The segments were then transcribed by native speakers, who also checked and made the necessary corrections if needed. The transcribers followed a transcription rules guideline document, which is shared along with the dataset.

The manual transcription task itself was carried out using Audacity[4] and the outputs of transcription are saved in a text file (with a .lab file extension, hence called a “lab” file). In the end, the transcriptions of the two channels are combined together to obtain a single lab file per call, as shown in Figure[5] 94% of all calls in ROXSD are manually transcribed.

The manual transcripts of non-English conversations were additionally translated into English using an automatic translation service. The outcome of the automatic translation was manually checked against errors and a similar lab file was gen-
Figure 2: Lab file structure for an example conversation.

Figure 3: YAML file structure for an example conversation.

Finally, the NLP Annotations category contains manual annotations of the English and German calls with respect to the following natural language processing (NLP) tasks:

**Named Entity Recognition (NER):** All person names, location names and times that are mentioned in the conversations are annotated. The location entity type includes geopolitical entities (countries, cities, states) as well as non-geopolitical entities such as hotels, streets, and restaurant names, among others. The time entity type includes absolute dates as well as times smaller than a day.

**Mention Disambiguation:** This task involves differentiating the annotated person names by assigning them to one of the parties (caller or receiver), or a third party who is mentioned in the call. This information helps the investigator understand the nicknames of the suspects being intercepted and their connections to other unknown persons/suspects. The third parties (mentions) are indicated in Figure 4 as nodes connected to the (real) characters with dotted green lines.

**Relation Extraction:** Relations from conversations are annotated. Currently, two types of relations are supported: the Current Location and Movement relations. The current location relation provides information about the approximate whereabouts of people or groups (e.g., “I am in Brno”). The movement relation tracks the movement of people from one location to another (e.g., “I am heading to Paris”).

**Topic Detection:** Each conversation is assigned to one of the following topics: Drugs, Work Conversations, Family-Friend Conversations, Money, Meeting, Other. The NLP annotations together with the transcription output are represented in a YAML file, as shown in Figure 3.

### 5. Image and Video Subset

In order to illustrate the interest of exploiting the visual data, ROXSD was complemented with images and videos representative of files which are likely to be found on a seized smartphone or computer, or simply downloaded from the Internet. These include videos of certain objects or locations which are recorded by individuals with their voices being heard in the background, and images in which multiple persons can be seen together. The captured images and videos enable the evaluation of face and scene matching technologies of ROXANNE to enrich the speaker network with additional nodes and edges. For instance, an edge is added between two speaker nodes when both persons are found (either through their voice or face) in the same video.

**5.1. Data Collection**

The ROXSD image and video subset contains three groups of documents originating from different sources.

The first group of videos consists of those recorded by the project partners. 15 individuals (14 of which are the same characters as in the calls subset) were asked to take video clips in 8 different outdoor locations while speaking. They were asked to mention some other characters’ names and the locations from the ROXSD story. Videos were captured with the participant’s own smartphone, resulting in a diversity of capturing devices. This group of files is aimed at evaluating the scene matching performances and the ability to match voices across devices (in particular, to match voices captured from tapped calls on one side and from videos captured from a smartphone on the other side).

The second group of videos was added to enable the evaluation of face matching technology, while respecting data protection regulations on processing biometric personal data. In practice, the SimSwap face-swapping framework\[11\] was adapted to consistently pseudonymize faces observed in the Realtime Selfie Video Stabilization Dataset by \[12\]. The SimSwap framework takes as input a video and a target face. It then consistently replaces the faces observed in the video by the target face. Faces generated by the StyleGAN2 model\[12\] were used as target faces, thus resulting in pseudonymized videos. A set of 47 identities was selected in the original dataset for which at least two different videos were available. For each identity, one video was then selected as a source for the enrolled face picture and the other(s) as videos to be processed (resulting in a total of 90 videos). To add complexity to the face matching process, the enrolled picture for each identity was selected as follows: first a face clusterization and cluster summarization process was run on the video, resulting in the selection of 5 representative face observations of the person; the first observation corresponding to the cluster’s face centroid, the second one to the face observation the furthest from the centroid (but still in the cluster). This second picture was selected as the enrolment picture for each identity. Figure 4 depicts an example of the two pictures and the corresponding video frames.

To constitute the third group, 4 characters of the ROXSD story were additionally asked to capture their selfie images and videos in 4 different locations in a setting that matches the ROXSD story, making up a total of 14 files. The faces involved in these images and videos were also pseudonymized.

---

1. https://github.com/neuralchen/SimSwap
5.2. Data Size and Statistics

There are a total of 154 videos in the first group of videos. In total, the images and videos contain information about 51 unique face identities, 15 unique voices and 12 unique locations. The total duration of videos is 1h 38min, with the following distribution of speech in 5 different languages: English (1h 8min 18sec), German (7min 24sec), Czech (1min 1sec), Greek (3min 33sec) and French (1min 6sec).

5.3. Metadata

A selection of 21 of the German speaking videos, and 23 of the English speaking videos were transcribed and annotated for NER. A ground truth file is provided along with each image or video, containing: (i) the list of speakers heard in the video, (ii) the list of faces observed in the image or video, and (iii) the list of scenes observed in the image or video.

6. Text Subset: ROXHOOD

The exploitation of publicly available social media sources (platforms, websites or forums) was a very challenging activity in terms of following the European and national legal, ethical, and human rights standards. One of the key challenges for data protection and ethics in social media landscapes is the usage of personal or sensitive data. In order to avoid any legal and ethical issues arising from accessing and working with publicly available social media datasets, the team decided to simulate the social media environment. To do so, a mock-up forum website was set up and a tool was created for collecting and indexing the available data within this environment. The result of this collection setup is the so-called ROXHOOD subset.

6.1. Data Collection

ROXHOOD was envisioned as a forum where legal and illegal content coexist, and where information about several topics is exchanged. It is based on Misago[14], a fully featured forum application developed to follow modern practices and trends currently used in web software development. It features complete moderation allowing admin-approved moderators to edit, move, hide, approve, delete, or close user posted content. This also includes the option to delete or block user accounts or avatars.

ROXHOOD is a fully featured internet forum solution where registered users are allowed to start a new thread, comment and perform searches as well as to set up private threads. The private threads feature allows users to create threads visible only to them and those other users whom they have invited. Registered users can write messages in text format, upload files of various formats (e.g. docx, jpeg, mpeg), include/provide links, comment on threads below the initial post, and mention each other (via @usernameX).

To collect data from ROXHOOD in a structured format, the ROXANNE crawler, RoCraw, was developed. RoCraw is a site-specific solution written in python which extracts data/info about the threads created in ROXHOOD and stores them in an Elasticsearch instance. The results can be exported as a JSON array, as shown in Figure 5. The majority of the partners created an account (user registration, which contained a fake nickname – not necessarily the same name as their character in the ROXSD story – and a fake e-mail address) and provided data

![Figure 5: ROXHOOD crawler JSON output.](image)

(6.2. Data and Metadata; Size and Statistics

A total number of 335 public posts by 48 registered users exists in the ROXHOOD subset. These posts are organised in 32 different threads. The ROXHOOD metadata is composed of a set of descriptions accompanying each forum post (message), such as thread and message identifiers, date and time of the message, flags defining whether the message contains a URL or an image, as well as the attachments of that message.

The collected dataset was processed using NLP techniques (including NER) to define metrics which describe the content and the type of entities. ROXHOOD includes mostly text messages and also some videos containing speech in English. Across 299 posts which were analysed, 110 entities are present with the following breakdown: 21 person(s), 49 location(s) and 40 time. Finally, regarding the content of English Videos, across 23 videos, 107 utterances and 91 entities have been defined. Again, for the entities of person, location, time the numbers are 44, 26 and 21, respectively.

7. Privacy and Ethics in Data Collection and Processing

LEA officers are frequently hesitant when it comes to the sharing of data, often due to security concerns. In cases where LEA officers are indeed willing to share data, there is a general lack of legal support to complete the necessary documentation (i.e., data protection impact assessments, as required for the processing of criminal data under Art.35, GDPR, and controller-processor contracts). The oversight of research activities requires specific expertise, which are very different from the usual activities that a LEA data protection or legal support officer might be asked to advise on. Lawful and ethical sharing of LEA data was conducted in other parts of the ROXANNE

---

[13] The primary legal framework applicable is the GDPR which replaced the EU Data Protection Directive.

[14] https://github.com/rafalp/Misago
8. Technologies Evaluated on ROXSD

A summary of different technologies developed on ROXSD is described in Table 2. The technologies are integrated as part of ROXANNE technological platform, a SW which is freely accessible for any LEA from European and academic partners recognized to be active in security-related research. It covers a wide range of applications, from speech and natural language processing to network analysis and video processing. Each technology employs specific methods and models tailored to its task, showcasing a diverse set of approaches and achieving promising results in various domains. In the field of speech, technologies such as Speaker Diarization, closed-set Speaker Identification, and open-set Speaker Identification utilize cutting-edge methods such as Voice Activity Detection, ResNet architecture, and x-vector-based approaches, achieving commendable results. Automatic Speech Recognition (ASR) for English and German utilizes models like XLSR-LFMMI and employs lattice rescoring techniques, showcasing competitive Word Error Rates and F1-scores. Language and Accent Identification leverage embeddings and probabilistic analysis for accurate linguistic characterization. In the field of Natural Language Processing (NLP), technologies involving cross-encoder methodologies and models like DistilRoBERTa demonstrate effectiveness in topic classification, while NER utilizes RoBERTa-large for precise entity identification. Network analysis technologies encompass diverse aspects, from mention networks with co-reference resolution and social influence analysis using Pagerank algorithms. Standout achievements include 100% accuracy in outlier detection and 75% accuracy in cross-network node matching. Within the video domain, Face Detection and Scene Characterization utilize cutting-edge techniques such as RetinaFace, ArcFace, and ResNet, showcasing high precision and recall. Collectively, these technologies highlight the versatility and efficacy of contemporary tools across varied applications, providing a glimpse into their specific methodologies and performance metrics.

9. Summary and Conclusion

ROXANNE project combined advances in speech, language and video technologies and criminal network analysis for the support of investigators in their daily work especially on large criminal cases, speeding up the investigative processes, as well as reducing the cost and burden to the society caused by organized crime activities. The ROXANNE Simulated Dataset (ROXSD) was designed and collected with the mindset to create a fictional but realistic case that resembles the data which could potentially be collected in a criminal investigation, and one which would enable the evaluation of the different technologies.
gies mentioned above. Its main benefit over existing datasets is its multimodal and multilingual structure, and the availability of extensive ground-truth annotations on the case, story, content and criminal network structure. The authors hope that these features will make ROXSD an interesting and valuable resource for the criminal research community in their increasing need for realistic datasets for their security-related research.

Acknowledgments
The authors would like to express their special thanks to the Police of the Czech Republic, Prague, and to Ralf Mösch (Head of Cybercrime Department) and Volker Brunnauer (Head of Data Analysis) from the Ulm Police, Germany, for their collaboration. The research described in this work was performed within ROXANNE – Real time netwOrk, teXt and speaker ANalytics for combating orgaNized crimE. This project has received funding from the European Union’s Horizon 2020 Work Programme for research and innovation 2018-2020, under grant agreement number 833635. Idiap team was partially supported by TRACY project[3] funded by Digital Europe Programme (DIGITAL) with Grant Agreement No. 101102641.

10. References