AUTHORS

Mirai Gendi (m.gendi@mail.utoronto.ca)
Periklis Andritsos (periklis.adritsos@utoronto.ca)

A Methodology for Event Log Generation from Unstructured Digital Forensics Data

A methodology combining NLP and Process Mining to transform unstructured forensic data into structured event logs, improving scalability, accuracy, and reproducibility in digital investigations.

AFFILIATIONS

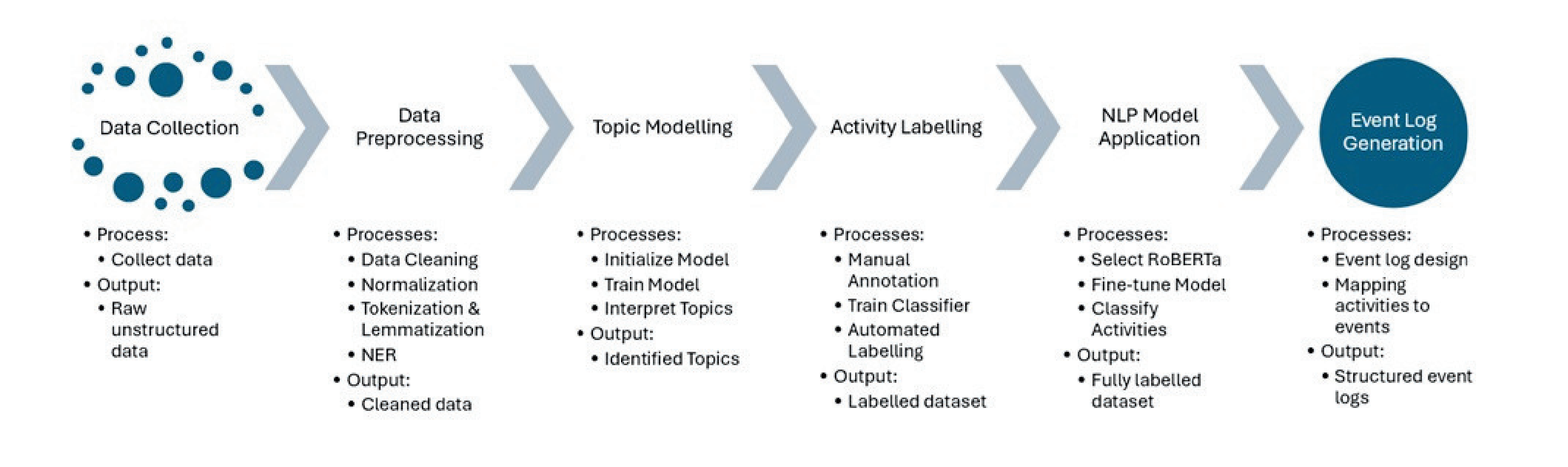
UNIVERSITY OF TORONTO

01. Introduction

Digital forensic investigations face challenges due to unstructured data from a variety of sources including social media, body-worn cameras, and multimedia files. Traditional methods struggle with processing such data, necessitating an automated approach. This study integrates Natural Language Processing (NLP) and Process Mining to convert unstructured forensic data into structured event logs.

03. Methodology

We collect unstructured incident reports from the Ontario Special Investigations Unit (SIU) archives and preprocess them (removing extraneous text, normalizing, and lemmatizing). A manually annotated subset defines digital-forensic activities for training a RoBERTa-based NLP model, while LDA topic modeling identifies latent themes. Named Entity Recognition extracts key entities, and the labelled text is converted into structured event logs with timestamps and metadata. Finally, a validation framework checks temporal consistency, merges duplicates, and cross-references the original case files for accuracy.



Data Collection
- Process:
  - Collect data
- Output:
  - Raw unstructured data

Data Preprocessing
- Processes:
  - Data Cleaning
  - Normalization
  - Tokenization & Lemmatization
  - NER
- Output:
  - Cleaned data

Topic Modelling
- Processes:
  - Initialize Model
  - Train Model
  - Interpret Topics
- Output:
  - Identified Topics

Activity Labelling
- Processes:
  - Manual Annotation
  - Train Classifier
  - Automated Labelling
- Output:
  - Labelled dataset

NLP Model Application
- Processes:
  - Select RoBERTa
  - Fine-tune Model
  - Classify Activities
- Output:
  - Fully labelled dataset

Event Log Generation
- Processes:
  - Event log design
  - Mapping activities to events
- Output:
  - Structured event logs

02. Objective

To develop an automated methodology using NLP and Process Mining to convert unstructured forensic data into structured event logs for scalable, accurate, and reproducible forensic analysis.

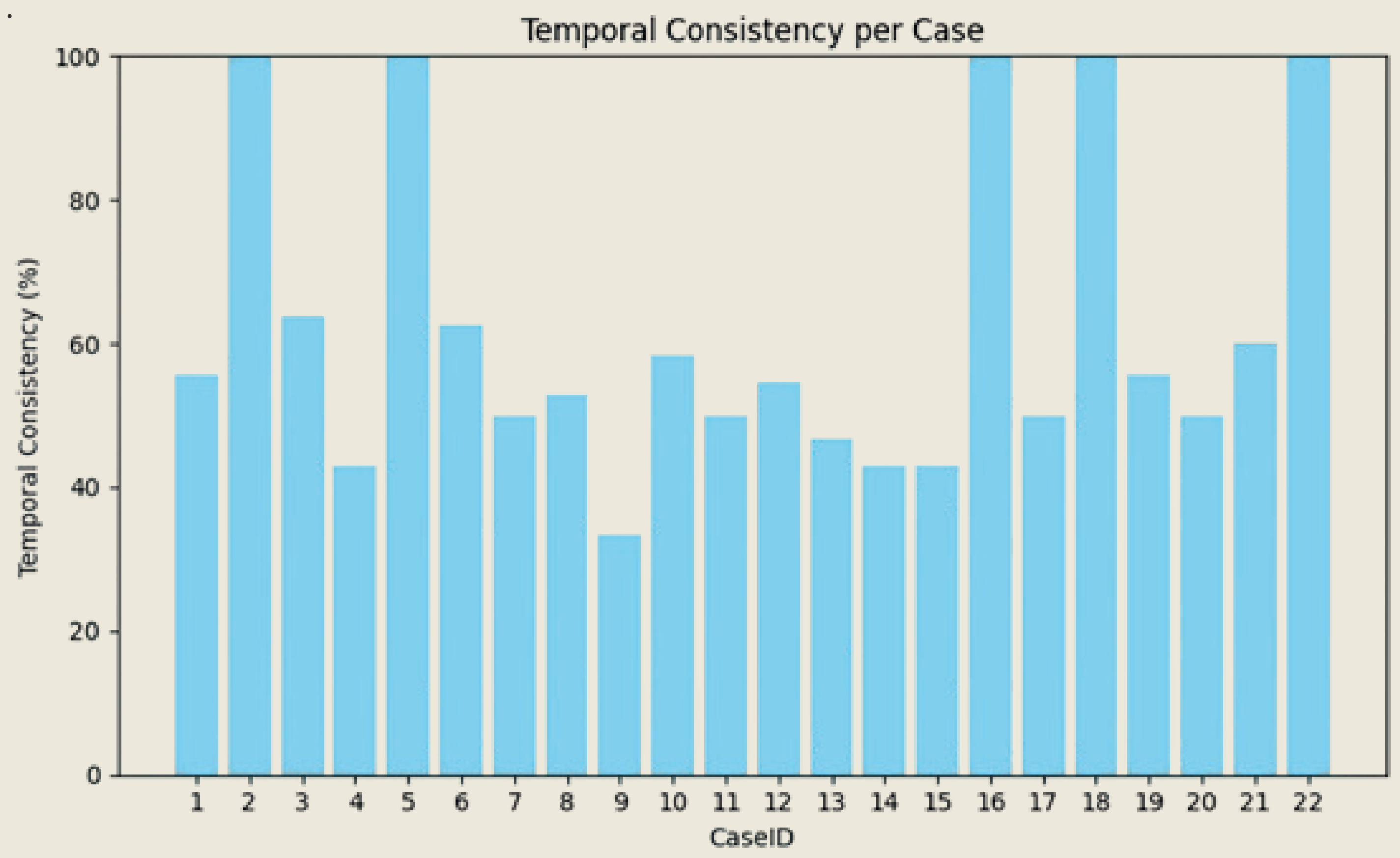
Related literature

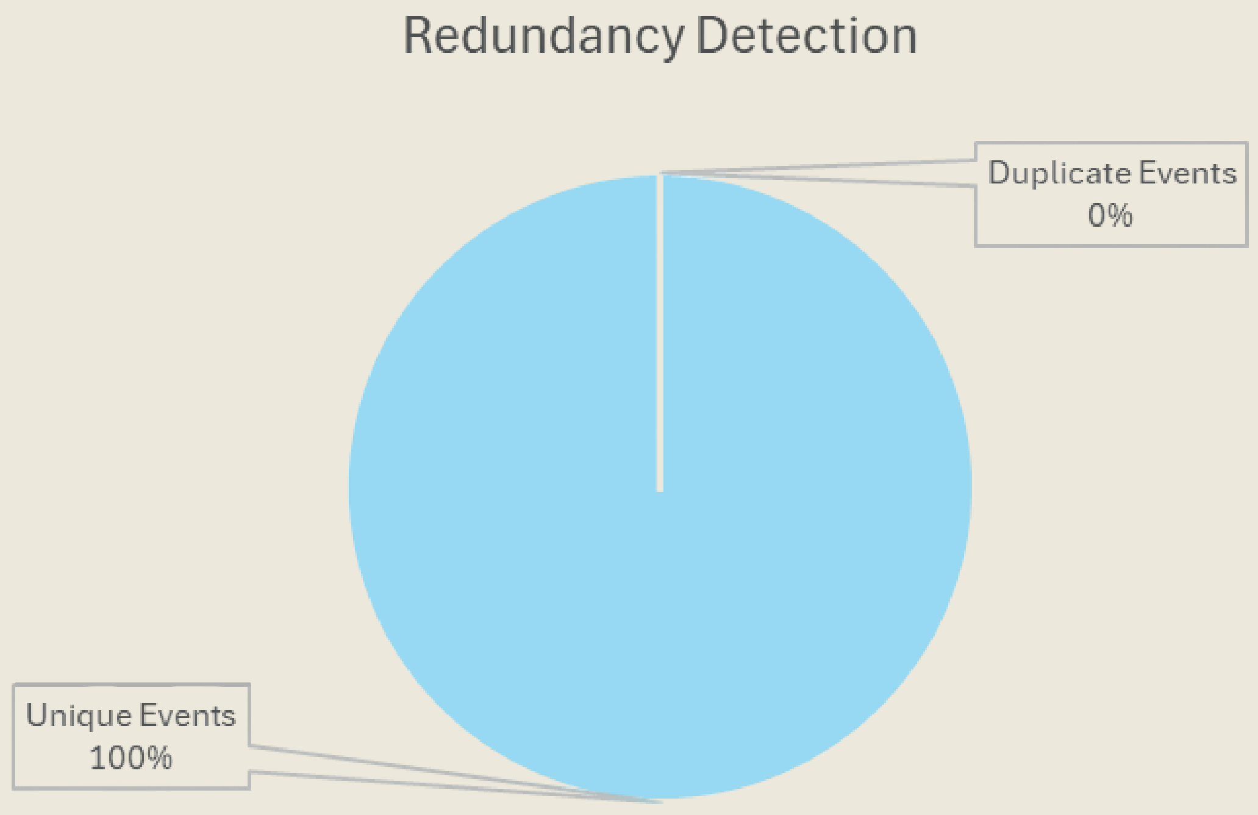
Prior work has applied NLP for text extraction and process mining for structured logs, but their integration for forensic event log generation remains underexplored.

04. Results/Findings

Our primary goal was to determine whether an NLP-driven methodology could effectively generate structured event logs from unstructured forensic documentation. The findings suggest that transformer-based models (e.g., RoBERTa) can provide accurate and scalable event classification, especially when combined with thorough data preprocessing and validation checks. Below are the key results:

- Accurate Classification:
  - A fine-tuned RoBERTa model reliably identified common forensic activities (e.g., "Forensic Imaging," "Data Extraction," "Report Generation").
  - NER extracted critical entities (people, locations, dates) and integrated them into the event logs.
- Structured Event Logs:
  - Automatically transformed text into time-stamped event entries (CSV format) to capture each activity and its timing.
  - Enables process mining for anomaly detection, audit-trail consolidation, and visualization of investigative steps.
- Validation Framework Success:
  - Temporal Consistency Checks ensured events were placed in proper chronological order.
  - Redundancy Detection merged overlapping or duplicate entries, minimizing clutter and preserving data integrity.





05. Analysis

We used a multi-step approach to analyze our text-based forensic data, focusing on NLP classification and event-log generation. The primary goal was to determine whether the methodology could accurately extract key forensic activities and represent them in structured event logs.

Data Preprocessing & Annotation

- *Applied cleaning, lemmatization, and a domain-specific dictionary to reduce noise.*
- *Manually annotated a small subset of documents to create ground-truth labels.*
- *Used these annotations to train and evaluate the NLP model.*

Classification & Confidence Scoring

- *Deployed a transformer-based model (RoBERTa) to classify sentences into activity categories.*
- *Adopted a probability threshold to flag uncertain classifications for manual review.*

Event Log Construction & Validation

- *Converted classified text into timestamped event logs with each entry tied to a specific time and activity.*
- *Performed temporal consistency checks to ensure chronological correctness and redundancy detection to merge duplicates.*

We automatically extracted timestamps and activity labels from unstructured text, allowing us to build a chronological event log

Each 'Finding' was identified via NER and text classification, then mapped to the relevant activity

Case ID	Timestamp	Activity	Action	Findings	Tools Used
1	2022-10-13 19:09	Evidence Acquisition	Interviewed Witness	Interviewed CW #1, CW #2, and CW #3	Interview Notes
2	2022-10-13 19:40	Forensic Imaging	Created forensic image	Forensic images of suspect's phone and laptop	Imaging Software
3	2022-10-13 20:00	Data Extraction	Extracted communication records	Extracted incriminating messages	Extraction Tool
4	2022-10-13 20:30	Data Recovery	Recovered deleted files	Retrieved deleted files from hard drive	Data Recovery Tool
5	2022-10-13 21:00	Report Generation	Created report	Report on metadata and communication logs	Reporting Software

06. Conclusion

This study demonstrates that transformer-based NLP can successfully extract and structure events from unstructured digital forensics data, yielding time-stamped logs that streamline investigations. Two key findings stand out:

- **Reduced Manual Effort**
  - The automated classification pipeline significantly decreases the amount of text needing human review, saving investigators' time and resources.
- **Increased Accuracy & Scalability**
  - By integrating domain-specific data cleaning and Named Entity Recognition, the methodology achieves more precise event detection and can handle large case files without overwhelming human analysts.

Limitations

- **Complex Data Types:** NER accuracy can drop when dealing with subword tokens or unique legal/technical jargon.
- **Computational Overheads:** Transformer models can be resource-intensive, potentially limiting their use in time-critical or hardware-constrained environments.

Future Work

- **Multimedia Integration:** Expand beyond text to handle body-cam footage transcripts or video metadata for more holistic forensics.
- **Enhanced Validation:** Refine validation checks to address new data sources and ensure reliability across diverse law enforcement workflows.