Forensic Technology
CENTER OF EXCELLENCE
A program of the National Institute of Justice

Sharing Knowledge | Advancing Technology | Addressing Challenges

IN-BRIEF SERIES (Part 3 of 3)

Digital Transformation of Cold Case Reviews: The Application of Text Analytics

Introduction to Report Series

Advances in forensic science capabilities represent opportunities and challenges for allied professionals involved in cold case violent crime investigations. Modern technologies can uncover important insights that may not have been previously possible; however, law enforcement and other associated agencies can struggle to efficiently leverage the large amounts of information associated with a cold case. Factors such as a lack of case file organization, decentralization of documents, and the time-intensive process of manually searching paper records for relevant details make it difficult to identify cases that may benefit from the application of new techniques and technologies (i.e., new compared to what was available at the time of the initial investigation or advances made to what was available at the time of the initial investigation).

> ***"We're constantly implementing new technology and go back to the cases to use it where old technology failed. We want to identify these cases without hand searching. With our high volumes of cases, we need to work smarter."***
>
> *—Carol Schweitzer, Supervisor, Forensic Services Unit, National Center for Missing and Exploited Children*

Objectives

- ► Introduce the concept of AI-based text analytics tools and their potential value to cold case reviews.
- ► Identify tools that can be leveraged to quickly draw important insights from information found in cold cases.
- ► Illustrate the realities of these emerging tools, which are being adopted by a growing number of industries.

The cold case review process is a collaborative effort during which professionals representing a variety of disciplinary backgrounds review all relevant case details to develop a strategy that will advance the investigation. The members of this multidisciplinary team (MDT), which may consist of law enforcement investigators, forensic science service providers (FSSPs), prosecutors, victim/family advocates, medicolegal death investigators, and sexual assault nurse examiners (SANEs), share common challenges within the case review process.

Tracking down and extracting relevant information to create an informed forensic or investigative strategy is both time and resource intensive. MDTs may look to technology to digitize paper cold case records, thus enabling multiple agencies on a secured network to search, share, and access files. The digitization of case files may support the future implementation of text analytics tools, which are enabled by artificial intelligence (AI) and can quickly identify key details, relationships, and patterns within or between case file records. Applying AI to the cold case review process can help MDTs become more agile and efficient in collaborating and developing valuable forensic and investigative strategies, but this digital transformation requires technical, time, and resource investments. This in-brief is the third of a three-part series that highlights the potential value, approaches, and considerations for digital transformation of cold case files and the case review process. Although created for cold case MDTs with an emphasis on the role of law enforcement and FSSP involvement, this series is valuable for all forensic science service providers involved in an MDT and criminal justice decision-makers. All three in-briefs focus on sexual assault and violent crime cold cases and the possible benefits of identifying evidence for additional or advanced DNA testing, but converting paper records into searchable electronic records may also lead to the identification and (re)testing of other types of physical evidence to advance other types of cases.[1] The information presented within these in-briefs can be applied to and benefit all cases regardless of the passage of time. **Specifically, this in-brief focuses on the value of text analytics and steps that MDTs may take to adopt these technologies.**

Key Takeaways

- Technology implementation may streamline the time-intensive process of cold case reviews, which is often shared across MDT members.
- In the future, investigators and FSSPs may employ text analytics tools, enabled by AI, to quickly uncover relevant case information. These tools can extract information from large text-based sources (e.g., narratives, case summaries, reports), such as names, types of evidence present, relationships between entities (e.g., identifying individuals who may be potential witnesses), and possible linkages across cases. Information gathered from these tools can increase the efficiency and objectivity of the case review process, inform forensic testing strategies, and identify commonalities that may not be readily apparent.
- Although off-the-shelf solutions do not currently exist for cold case investigations, these tools are being adopted in other industries such as healthcare. Text analytics tools that are commercially available may be configured to fit the organization's needs, with the help of technical experts. Organizations may also work with technical experts to create a software tool to address these needs.
- With the help of research funding or corporate support, some organizations (e.g., the Los Angeles Police Department and the National Center for Missing and Exploited Children) have implemented text analytics or have begun the process of creating text analytics tools for analyzing case file data. These tools were either developed by technical experts or heavily configured by a vendor.
- MDTs looking to implement these tools in the future should consider the time, resource, and technical investments needed to create these tools. To realize the full benefits of AI, organizations must train the tool and invest resources to maintain these tools.
- Commonly expressed needs for these tools and their high development costs are an impetus to subsidize development of a text analytics tool that could become an open-source software. This would require a significant up-front investment but could eliminate the need for multiple organizations to develop the same type of tool.
- Digitizing case files is a necessary first step for organizations to implement text analytics tools (see the second in-brief of this three-part series, "***Digital Transformation of Cold Case Reviews: Digitization of Case Files***").

Context

Reviewing key details of the initial investigation is an important first step of a cold case review.

Ongoing forensic science improvements have enhanced FSSPs' and law enforcement agencies' abilities to obtain just resolutions for victims and co-victims (e.g., family members) impacted by crime. Developments in techniques, technologies, databases, and forensic testing methods may help resolve active casework and cases that have gone cold.

In collaboration with an MDT, investigators may employ a systematic review process to develop an investigative strategy that advances the case toward resolution (e.g., identifying evidentiary items that may benefit from (re)testing). This cold case review process is a cross-collaborative effort that ensures all case details and case status are known across the MDT. During this process, members of the MDT work to identify:

- **Basic case information.** Who are the individuals in this case, including victim(s), suspect(s), and witnesses? What type of crime occurred? What locations and other details may be important to the case?
- **Case status.** What evidence is available, and where is it stored? Who is available to (re)interview about the case? What investigative leads need follow-up? When was the last time this case was reviewed?
- **Opportunities for advanced forensic testing.** Is there remaining evidence that would benefit from (re)testing using advanced forensic techniques and technologies?
- **Case linkages.** Are there similar names, details, potential evidence, or locations across multiple case files?

For this in-brief series, the FTCoE has defined the following terms:

- **Cold Case:** Cases where investigative leads have been exhausted with specific focus on cases related to violent crimes, including homicide, sexually motivated homicide, and sexual assault.
- **Resolution:** An outcome resulting from a full investigation in which a case is cleared by arrest (i.e., an individual has been arrested/charged with the crime or the case is turned over to prosecution) or closed as a result of a circumstance outside of the investigating agency's control that prevents an arrest, charge, or prosecutorial action against the individual (e.g., individual is deceased).[2]

More information about the cold case review process and considerations for forming an MDT can be found in the first in-brief of this three-part series, "***Digital Transformation of Cold Case Reviews: Prevalence, Challenges, and Benefits of Just Resolutions***."

Cold case review processes are complicated because of the large volume of information that criminal justice professionals must consider.

During the cold case review process, criminal justice professionals typically manually search case files and compile relevant information. The MDT reviews this information to determine the current case status (e.g., identifying evidence that has previously been submitted for forensic testing and any remaining evidence that is suitable for (re)testing).

This can be labor intensive and lead to inefficiencies in the cold case review process. Case files include documents in various formats and can be very large—in excess of 1,000 pages. Often, paper files lack organizational structure that makes it difficult to locate key documents, such as laboratory reports, requiring the MDT to manually search through the entire file for this information. This manual process of reviewing and extracting important details can take multiple hours per case for an experienced investigator and longer (up to a week) for an intern. The inefficiency of paper files causes MDTs to dedicate a disproportionate amount of time toward gathering case information and away from developing investigative strategies. **Although MDT members provide varied perspectives to a cold case review, they**

share the need to quickly find relevant data from case files that can help them understand key details about a case, identify patterns, and inform next steps, such as forensic testing.

Text Analytics Tools and the Potential Value to Cold Case Reviews

Text analytics is the practice of analyzing large amounts of text-based information to uncover important insights. The inputs for text analytics tools are usually unstructured data, where the text data are not organized in a predefined manner (which may include text-heavy narratives or long pages of text that are not organized in forms).[3]

Text analytics tools are enabled by AI, a discipline that allows computer analyses to "perceive and respond independently and perform tasks that would typically require human intelligence and decision-making processes, but without direct human intervention."[4] More specifically, these tools leverage natural language processing (NLP), a subset of AI that relates to a computer's ability to understand and generate human language as it is spoken or written. ***Exhibit 1*** demonstrates how text analytics tools relate to AI and NLP. NLP is present in many software programs on mobile devices or computers, including programs such as autocorrect and autocomplete, search engines, and chatbots. With NLP, these programs are configured or trained to make sense of and "understand" human language inputs and react in some way (e.g., automatically extract insights, suggest alternative spelling, complete a word being typed out). These tools can extract important information from "unstructured" text data (e.g., a witness statement or case summary).

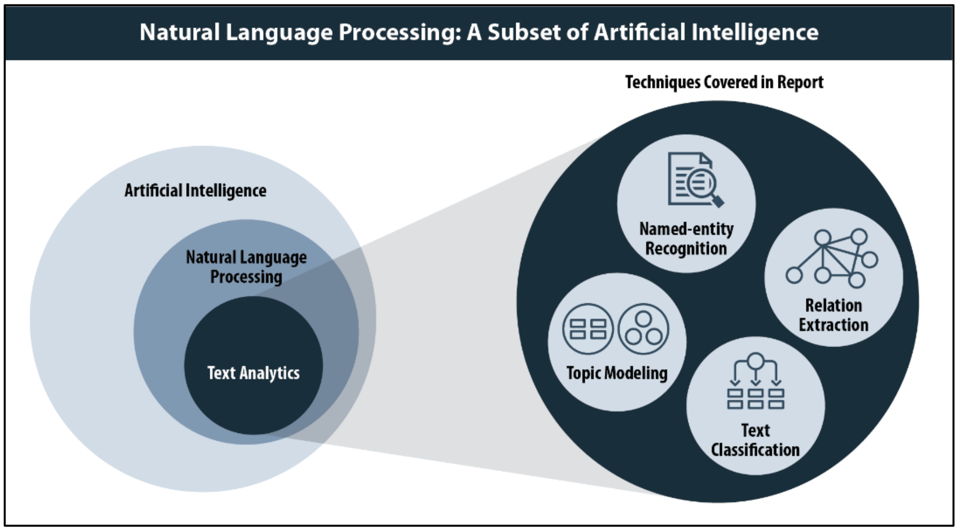

Natural Language Processing: A Subset of Artificial Intelligence
Techniques Covered in Report
Artificial Intelligence
Natural Language Processing
Text Analytics
Named-entity Recognition
Relation Extraction
Topic Modeling
Text Classification

Exhibit 1. Text analytics tools leverage techniques within the NLP discipline (a subset of AI) to help draw insights from large text-based files.

Text analytics tools that are commercially available and adopted by multiple industries may perform key tasks that can help drive timely and just cold case reviews.

Analyzing large, cumbersome volumes of information is not unique to the criminal justice community—many industries today (e.g., healthcare and e-commerce) use AI to derive insights from large and sometimes disparate datasets. Tools leveraged by these industries may be extended to the criminal justice community to help MDTs quickly find information relevant to a cold case investigation. These tools can perform tasks such as the following:

- **Automatic extraction of case details that may be of interest**. Tools can scan text information and identify names, places, and keywords that may be relevant to the investigation. These tools can also direct the user to this information within the file.
- **Searching capabilities that inform forensic testing strategies**. MDTs can leverage keywords and phrases to identify testing strategies that have been employed and future testing opportunities. Additionally, the presence—or absence—of terms can code or tag specific cases, such as sexually motivated homicides or sexual assaults.

- **Identification of common information that suggests links across cases.** By capturing potential relationships between keywords (e.g., witnesses and locations) and commonalities across cases (e.g., case details such as a "white van"), text analytics tools can suggest potentially informative linkages.

Implementation of AI-based text analytics tools may streamline and reduce bias in the cold case review process.

Although tools cannot replace the experience and skills of the MDT, text analytics tools can effectively augment and streamline key tasks in the cold case review process. In addition, these tools can support national search programs, such as the Federal Bureau of Investigation's National Incident-Based Reporting System and the Violent Criminal Apprehension Program, which rely on agency reporting to create comprehensive, up-to-date, and meaningful databases. These databases may also inform other tools like the Murder Accountability Project's dashboards, which can identify potential hotspots or connections between crimes. In a case review, these tools can:

- **Identify additional testing opportunities for FSSPs**. These tools can identify partially tested sexual assault kits or inconclusive results.
- **Reduce time needed for document searching**. These tools can automatically search against predefined fields and can support custom queries across multiple cases allowing agencies to spend more time identifying appropriate testing strategies and less time searching case files.
- **Provide insights across cases**. Some tools can outline potential linkages and commonalities across a wide variety of cases, suggesting connections that may not be apparent without high-level aggregation and analysis.
- **Provide value for active and ongoing cases, in addition to cold cases.** The tools and approaches discussed in this in-brief, designed to uncover important details from large amount of cold case information in a timely manner, can be used to connect and search across active cases.
- **Facilitate objective investigations.** These tools can aid in the elimination of human factors, biases, and subjectivity allowing investigators to be more objective and less influenced by non-critical or non-relevant information.

Several text analytics techniques can pull case information that the MDT manually gathers in a case review.

Text analytics tools may leverage one or more techniques to uncover relevant case information. ***Exhibit 2*** provides an overview of some techniques that are especially useful for identifying important information in a cold case review process. More information on these techniques can be found in the appendix.[a] AI-enabled text analytics tools can be diverse and may leverage one or more of these techniques to gather information relevant to a cold case.

Technique	Capabilities	Impact to Cold Case Review
Named-Entity Recognition (NER)	Identifies and categorizes entities (e.g., names, locations, domain-specific terms) into a particular category	These tools can automatically search for important details of the case. For example, users may use the tool to search for potential evidence for DNA testing using the terms "blood," "semen," "urine," "PSA," and "serology." Users may also search for evidence that may have been kept ("retained" alongside "evidence") or instances when retesting might be appropriate ("insufficient" or "inconclusive").

[a] Although this report outlines AI-based approaches that may be useful for cold case file review, the approaches covered are not exhaustive.

Technique	Capabilities	Impact to Cold Case Review
Relation Extraction	Identifies potential linkages or relationships between words in a sentence (e.g., Person A is a witness or known associate)	These tools can help users understand how different entities (information such as people and places) are related to each other, based on the context of a sentence. Used with NER, it can perform link analysis, which visually maps relationships across one or more cases. These links can provide an investigator or FSSP with an overview of important case details and present patterns (e.g., similar crimes happening in a particular area).
Topic Modeling	Scans a text file for similar word groups and expressions, and clusters them based on shared characteristics	These tools can identify details that are common across two or more cases that may be difficult to observe on a case-by-case basis. (e.g., "white van" is consistently appearing in multiple cold cases).
Text Classification	Sorts documents or text into specific taxonomies and categories outlined by the user; can identify specific clusters of words or expressions	These tools can help code, label, or sort case files based on details in the text. For example, identifying all documents that might be (or be closely related to) laboratory reports, based on the presence of different words or concepts.

Exhibit 2. AI techniques leveraged by text analytics tools.

Because a "perfect" off-the-shelf tool currently does not exist, MDTs may lean on several approaches to implement text analytics tools, though they all require varying technical investments.

Although off-the-shelf tools have not been created specifically for cold case review applications, MDT members may consider some products or pathways to create a tool that addresses their needs. Agencies may consider:

- **Applying law enforcement–specific tools.** Some investigative case management, intelligence, and threat monitoring tools use text analytics to search through documents and make connections across cases.
- **Purchasing general off-the-shelf systems and configure them to fit their needs.** Off-the-shelf general text analytics tools can be customized to cold case applications through training and technical support. These off-the-shelf systems offer a combination of flexibility and structure.
- **Building a custom system using technical support that directly addresses cold case needs.** Contracted developers with experience in data analytics can work with FSSPs and agencies to build a system that specifically addresses their needs, or units with access to programming talent can build a system from open-source software products, which are freely available to the public for their use.

Exhibit 3 provides examples of approaches (and specific product and service examples) that MDTs may consider if implementing text analytics in their workflows.

Approach	Description	Product Examples
Applying law enforcement–focused tools	**Investigative management and intelligence platforms:** Platforms to manage ongoing, active investigations or those that analyze and react to potential safety threats, pull insights from a large amount of text-based data, and may be applied to cold case reviews. These tools have domain-specific language that can pull out entities and relationships important to a cold case. Many of these tools have advanced searching capabilities, allowing investigators to search for specific concepts, or for specific entity types (e.g., names, locations).	Intelligence platforms, such as Sintelix, may leverage NER to pull out insights from unstructured documents. They may offer machine-learning capabilities to add new entities and train the tools. Siren Platform has developed an an integrated investigative intelligence platform where investigators can take data from multiple sources and identify relationships and patterns through entity recognition and link analysis.
Configuring general text analytics tools	**Standalone text analytics platforms:** Web-based analytics platforms can range significantly in complexity. Standalone sites allow users to apply different text analytics techniques including NER, relation extraction, topic modeling, and text classification, to a data set. These tools may enable training via datasets and are often delivered as application program interfaces (APIs), which allow two software programs to interact with each other. These tools typically have an interface that is easy to use by non-technical experts.	NetOwl offers an extractor tool that can identify over 100 different types of entities, many of which are specific to law enforcement and laboratories and can be customized. Although this tool is pre-trained to identify some entities and relationships of importance to law enforcement, it does not create visualizations, requiring users to interpret these "raw" data outputs. Rosette Analytics offers both on-premise and cloud-based deployments of their product suite. These products conduct specific text analytics techniques and visualization of the data outputs. Monkey Learn helps users extract specific keywords from a text-based dataset and can be used to build custom extractors and classifiers. The data pulled from the unstructured data (such as a case narrative) can be visualized and analyzed for patterns.
Configuring general text analytics tools	**Text analytics platforms integrated into cloud services:** Similar to standalone web-based analytics platforms, these platforms allow users to extract important data with text analytics techniques. These tools are a part of a greater suite of tools related to data storage in the cloud. Organizations already storing their data in these cloud services could apply these NLP techniques to their datasets. These off-the-shelf systems offer a combination of flexibility and structure and can be configured to MDT needs (usually through training to recognize domain-specific language, such as types of evidence). These tools typically have an interface that is easy to use by non-technical experts.	Amazon offers Comprehend, a Criminal Justice Information Services–compliant NLP tool that can be trained for recognition of specific entity types and other text analytics tools. Google Cloud Natural Language AI, Microsoft Azure Cognitive Service for Language, and IBM Watson Natural Language Understanding also offer a variety of Cloud APIs that can perform several NLP functions, visualization, and data storage.

Approach	Description	Product Examples
Building custom system tools for a bespoke system	**Organizations with capabilities to build software products:** Agencies may leverage data science consultants or consulting firms that can create customized, integrated solutions that work for their needs. These tools often enable seamless data integration because they are built specifically for the needs of the client. However, these tools are expensive to build and maintain.	Consulting firms with strong data analytics capabilities, such as Booz Allen Hamilton, KPMG, and Accenture have created custom analytics tools for agencies. There are some smaller firms that focus specifically on the creation of NLP tools that could address agency needs. For example, Basis Technology can create custom extraction models for entities and relationships that are significant. Agencies may consider leveraging university researchers to help build relevant tools.
	Open-source NLP toolkits: These free resources can perform a variety of complex NLP techniques and may offer datasets to help train the tools. Most of these tools require proficiency in a programming language (e.g., Python) and technical experience to easily use. Outputs from these tools may be difficult to interpret by a non-technical expert.	spaCy and Natural Language Toolkit are two well-known examples of open-source NLP toolkits. These tools can operate in Java, Python, or other programming languages and can use data from network storage, cloud-based datasets, or other sources.

Exhibit 3. Potential approaches to implementing text analytics for cold case reviews.

Current Implementation of Text Analytics in the Criminal Justice Community

Widespread implementation of AI-enabled text analytics into FSSPs or law enforcement agencies is currently limited by technical and resource realities.

Although AI-enabled text analytics tools are being implemented in different industries, adoption for cold case investigations is low. MDTs looking to transform their cold case review process with AI should consider the following:

- **Customization and training investments are directly linked to the value of the tool**. AI tools must be trained to recognize key words and phrases of value to an MDT. Some off-the-shelf tools have been pre-trained or are able to recognize entities or patterns based on simple rules (e.g., whether a word is capitalized or where it fits within a sentence) but may not be able to identify entities or linkages accurately enough to save time for users. With help from technical experts, users can train the tool (see ***Appendix A: NLP Techniques that May Support Cold Case Data Analysis***) with real data to enable the tool to pull out valuable information from a dataset. Although training the tool increases setup and implementation time and costs, it will ultimately translate to a more accurate tool.
- **Technical investments for creating a value-adding solution may exceed its value for a singular organization.** Because off-the-shelf tools cannot easily address all cold case investigation needs, agencies must configure available software programs or build them from scratch. Technical labor needed to create and train these systems is the most significant investment for building these tools. Depending on project needs and timeline, labor costs can be upwards of $50,000 and exclude software, hardware, server costs, and ongoing maintenance. These costs are often prohibitive without external funding.

Sharing Knowledge | Advancing Technology | Addressing Challenges

Despite the limited adoption of tools, several trailblazing organizations are working to overcome these barriers and realize the potential of AI for cold case investigations.

Known use of these tools for cold case investigations is limited, but ongoing research and increasing digitization efforts from MDTs increases the potential for an AI-enabled future. The following section summarizes efforts to develop AI-enabled text analytics tools that can be leveraged for cold case applications and research efforts to help pave the way for future analytics tools.

The National Center for Missing and Exploited Children (NCMEC) has considered data mining tools for case identification and linking, key word searches, and evidence searches.

NCMEC assists the criminal justice community with long-term investigations, including cold case investigations involving child victims, by providing case consultations and other resources.

NCMEC has used emerging AI-based technology to analyze large amounts of data. Since 2010, they have partnered with data mining and analytics firm Palantir to help them sift through their CyberTipLine alerts (which are structured forms) and note patterns between reports that may suggest an investigative lead. This collaboration with Palantir has focused extensively on real-time tips to extrapolate important statistics and trends. Palantir's technology saves significant time by finding linkages across tips, which law enforcement uses to respond more efficiently to time-critical situations.[5]

However, NCMEC is not using the technology for cold case reviews; today, personnel must review entire case files to glean relevant information. In 2017, prompted by an office move, NCMEC digitized 8,000 case files to transfer their large volume of paperwork to an electronically secure network. NCMEC recognizes case file digitization as a step toward future adoption of AI-enabled tools to identify connections between cases. NCMEC pinpointed the value that AI could bring to reviewing case files:

- **Case Identification**. The ability to identify cases that meet grant criteria and can be worked using the funds.
- **Link analysis**. A visual means of connecting important data within and across cases.
- **Video Analysis.** The ability to analyze videos for common characteristics, such as backgrounds, locations, types of clothing or vehicles, which may generate additional investigative leads.
- **Key Word Search.** The process of interpreting statements with sentiment analysis or searching for key words.
- **Evidence Search.** The capability to search laboratory reports to pull out evidence and test results, potentially assisting evidence retesting.

The Los Angeles Police Department (LAPD), in collaboration with Justice & Security Strategies, Inc. and the University of California Los Angeles, is developing an investigative tool from LAPD's homicide data.

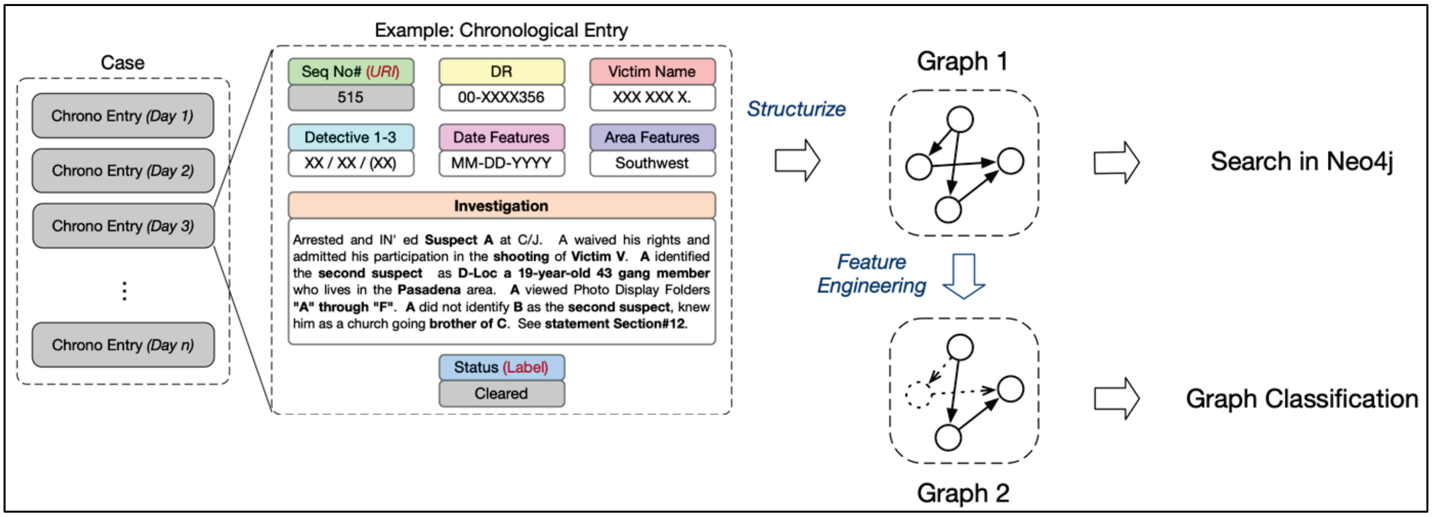
The LAPD, one of the country's largest police agencies, developed a practice to systematically document homicide investigations in the 1980s, which led to a large body of standardized case information. For each homicide investigation, the department organized case file data into standardized binders referred to as murder books. Each of these binders is divided into 26 sections and can contain anywhere from 800 to 1,000 pages of information. The sections are organized by tabs and contain a variety of forms, files, notes, and written entries. Roughly 4,000 of these case file binders have been digitized, with manual data entry completed for 3,309 of these binders as of February 2022.

The standardization and quantity of these binders positioned the LAPD's case files as an ideal dataset for research on the use of AI tools for case investigations. For this research, the LAPD partnered with consulting firm Justice & Security

Strategies, led by President Dr. Craig Uchida, who partnered with a University of California Los Angeles team led by Dr. Jeff Brantingham. Drs. Uchida and Brantingham evaluated the factors involved in resolving cases, with the goal of creating an investigative tool based on machine-learning methods. This partnership was supported by funding awarded to Justice & Security Strategies (JSS) from the National Institute of Justice (NIJ) (2018-75-CX-0003) and to LAPD from the Bureau of Justice Assistance (BJA) (2018-WY-BX-0002).

First, JSS focused on the variables of homicide investigations that appeared to be essential in resolving these cases. Using the digitized case files, they systematically coded (or tagged by category) variables from each case file, such as victims, suspects, and types of evidence (e.g., sexual assault kit, firearm). JSS then ran advanced statistical analysis on the coded data to identify factors important to case resolution. The research team used the digital case files to apply machine learning and build "knowledge graphs." Knowledge graphs extract information from the case file and show potential links (e.g., key individuals and how they may be associated to specific locations). To build the knowledge graphs, the team used AI techniques, such as NER and relation extraction, to analyze chronological entries (time/date stamped entries in the case record) made by investigators. These "chrono entries" were brief descriptions of steps taken by detectives during the investigation, such as an interview with a witness or record of a recovered weapon. Each case can have hundreds of chrono entries, depending on its complexity (see ***Exhibit 4*** for a visualization of this process).
The knowledge graphs can be used by LAPD to develop the following tools:

1. **An internal search engine that can pull related case information (i.e., crimes at a particular location).** This is similar to the way Google or Wikipedia organize, search for, and produce information in search results.
2. **An investigative tool to predict the likelihood of solving future cases.**



Seq No# (URI): 515
DR: 00-XXXX356
Victim Name: XXX XXX X.
Detective 1-3: XX / XX / (XX)
Date Features: MM-DD-YYYY
Area Features: Southwest

Investigation

Arrested and IN' ed **Suspect A** at C/J. A waived his rights and admitted his participation in the **shooting** of **Victim V**. **A** identified the **second suspect** as **D-Loc a 19-year-old 43 gang member** who lives in the **Pasadena** area. **A** viewed Photo Display Folders **"A" through "F"**. **A** did not identify **B** as the **second suspect**, knew him as a church going **brother of C**. See **statement Section#12.**

Status (Label): Cleared

Exhibit 4. Example of knowledge graphs built from chronological entries using text analytics techniques. Graciously provided by JSS (supported by NIJ award 2018-75-CX-0003 and BJA award 2018-WY-BX-0002).

After the digitized case files have been systematically coded and JSS has completed a working model of their investigative model built on the knowledge graphs, the team will apply the model to identify 100 cold cases with a higher likelihood of solvability. LAPD will reinvestigate these cases to test the model.

The Dutch Police Force is the only known agency that has expressed intent to develop AI tools specific for cold case investigations.

In 2016, the Dutch Police Force founded Q, an innovation division to resolve persistent challenges with novel solutions. Through short-term exploration of innovative approaches where staff can participate and learn about new investigative methodologies, the team focused on cultural change within their agency. Q is currently developing an AI-based machine to read cold cases to decide which "contain promising evidence that could lead to" case resolution.[6] The AI model will rank cases based on their solvability, helping the Dutch Police Force effectively dedicate resources to cold cases. This technique is relatively new, and an evaluation of the efficacy of the tool will not be available for some years.

Several federal funding efforts have supported research and development efforts for law enforcement tools that leverage AI.

Although these awards do not directly support development of tools to streamline cold case reviews, they support law enforcement tools that leverage text analytics methods to gather important insights from a large amount of text data.

- Using Sentiment Analysis and Topic Modeling in Assessing the Impact of Police Signaling on Investigative and Prosecutorial Outcomes in Sexual Assault Reports (2018-VA-CX-0002)

This Case Western University project investigated how a responding officer's language used in a sexual assault report impacted the investigating officer's process. This project used NLP technology—more specifically, topic modeling and sentiment analysis—to evaluate over 6,300 written narratives in police reports of sexual assaults.

- Cyberstalking - Research and Evaluation to Enhance Criminal Justice (2020-R2-CX-0002)

Using NLP, RAND Corporation leveraged text information from criminal cyberstalking cases to understand strategies used to target cyberstalking offenders and the outcomes of these prosecutions. The goal of the project was to identify probative factors that may lead to successful cyberstalking prosecutions.

Development Pathways: Enabling Future Implementation of Text Analytics for Cold Case Applications

Implementing text analytics tools for cold case review may help MDTs direct more time to analyzing and making connections and less time to manually searching for information. Although commercial products dedicated to cold case investigations do not exist, organizations should consider the following to (1) work toward implementation of future products or (2) configure or build products that fit their needs.

MDTs should understand organizational needs and goals for adopting tools for cold case reviews.

Implementing tools requires buy-in from multiple perspectives in an organization, including personnel directly involved in cold case reviews and agency leadership, information technology personnel, and other supporting MDT members. These organizations should ask themselves:

- **Is there a need for digital transformation of cold case files beyond digitizing?** By converting paper text records into a machine-readable form, MDTs can search for specific keywords or phrases. Digitizing and AI-enabled text analytics tools both drive toward more efficient cold case reviews, but the costs of implementing analytics tools is significantly higher than digitization costs.

- **What are the goals for implementing text analytics?** Does the MDT want to streamline investigations, improve searching capabilities, and enhance ability to draw links between different cases? Does the MDT have a consensus on successful use of text analytics tools?
- **What sort of leads could this tool help identify?** MDTs involved in the cold case review processes are encouraged to work together to establish a lexicon of keywords and phrases they typically look for in case documents. These could be terms that suggest evidence has been tested or that a type of evidence is present (e.g., biological fluids). Text analytics tools are configured specifically around these lists.
- **How should information gathered from the tools be used?** Text analytics tools can point to helpful information, though their outputs do not tell the entire story. MDTs should not make decisions solely on outputs of these tools; rather, these outputs can be a starting point for further inspection of a case file. Similarly, MDTs should value and investigate other appropriate leads that are discovered without the help of an AI tool.

Consider required organizational and technical investments.

MDTs must consider their access to labor, technical capital, and budget when determining a strategy to implement text analytics. Implementation of these tools, like many other tools, requires time and labor investments to set up, test, and ultimately train new users, and MDTs should consider these before purchasing a product or service. MDTs should understand the following:

- **Digitizing is a pre-requisite to adopting text analytics tools.** Converting case file text to machine-readable text is a necessary first step to preparing information for analysis tools. This step may require significant efforts to aggregate, sort, and systematically digitize the information. More information about digitizing can be found in the second in-brief of this three-part series, “***Digital Transformation of Cold Case Reviews: Digitizing Case Files***.”
- **Tools—regardless of approach—will likely need up-front and continuous training**. For a text analytics tool to search for and return results that align well with a specific industry or end user need, it must be trained to recognize and pull this information. Several off-the-shelf tools can recognize general concepts but may pull irrelevant information or miss important terms that are relevant to the criminal justice profession. While considering the goals for implementing these tools, future users must recognize that an accurate, effective tool will most likely require some level of training. This means that developers would need clear, labeled examples of information that criminal justice professionals look for in a cold case review.
- **Consider what existing systems, funding sources, and technical staff may serve as a foundation for these tools.** Implementing these tools will require a combination of labor and software to select, integrate, deploy, train, and maintain the system. Agencies may consider leveraging existing information technology experts or data science consultants and contractors to help them configure a system that works for the agency’s needs. Laboratories and agencies should also consider funding mechanisms from NIJ (e.g., Research and Evaluation in Publicly Funded Forensic Crime Laboratories) and BJA (e.g., Paul Coverdell Forensic Science Improvement Grants) to help support the implementation of tools.

Does your agency have many handwritten files?

Documents with handwritten information add another layer of complexity to digitizing case files. Current technology can easily convert typed text to a machine-readable format that can be imported into text analytics tools but cannot accurately convert handwritten text. These documents must be manually transcribed.

Lean on the technical expertise of vendors and previous adopters.
Text analytics is a rapidly growing field with early adopters across many industries and companies starting to create products for specific end users. Communicating cold case review needs to product and service vendors can help inform MDTs of emerging products that fit their needs (and those that do not) and help signal the need for an off-the-shelf product. Early adopters of the tools, even outside of the criminal justice community, can share the opportunities and realities of these tools.

Talk to product and service vendors before implementing any solutions. Although the challenge of making sense of extensive, disparate data is shared across many agencies and laboratories investigating cold cases, individual needs could vary extensively. There are no off-the-shelf solutions, but several vendors may be able to configure or create a system that meets these needs. Organizations considering purchasing a system will benefit from speaking with multiple vendors to communicate their needs and constraints and schedule demonstrations of their product. The following are questions an agency may ask a vendor:

1. Does your platform have the technical capabilities to address our needs today? What should be configured to meet our needs?
2. What is a reasonable timeline for creating/implementing this tool? Based on our needs, what sort of training or up-front work is required prior to launch?
3. How does this tool impact MDT workflows, and what features can streamline searching? (e.g., can the tool point to where certain keywords were present in a case file?)
4. What do the outputs of these tools look like? Is it easy for a non-technical expert to interpret?
5. What support do you offer for training?
6. What technical specifications or additional hardware/software is needed for successful implementation?

Identify and learn from applications for AI-enabled text analytics outside of the criminal justice community. Use of text analytics tools is currently limited in law enforcement, but MDTs can learn from industries that are actively adopting tools for application to large amounts of unstructured data. The healthcare industry, for example, is beginning to leverage AI, not only for clinical automation to increase efficiencies but also for early disease detection and predictive diagnostics. Like law enforcement and FSSPs, they are working with unstructured data that have a specific lexicon important to the community. Companies have created analytics tools trained based on real patient datasets to understand and analyze complex medical jargon.

Conclusion

AI-enabled text analytics is a rapidly evolving field, with significant implications for the criminal justice community. No technology will ever replace careful MDT analysis, but text analytics tools have the power to transform how MDTs review cold cases. With successful implementation of text analytics tools, MDTs may be able to spend more of their time in deep analysis and investigation with less time spent manually searching decentralized paper case files. Rapidly increasing adoption of these tools across industries and research efforts specifically related to adoption of AI tools in the criminal justice community signal the development of tools that will address MDT needs for cold case review processes. Organizations can prepare for tomorrow’s advanced text analytics tools by digitizing their case files, gaining consensus on appropriate focus areas and use of the tool, and learning from external industries and vendors as the technology continues to develop.

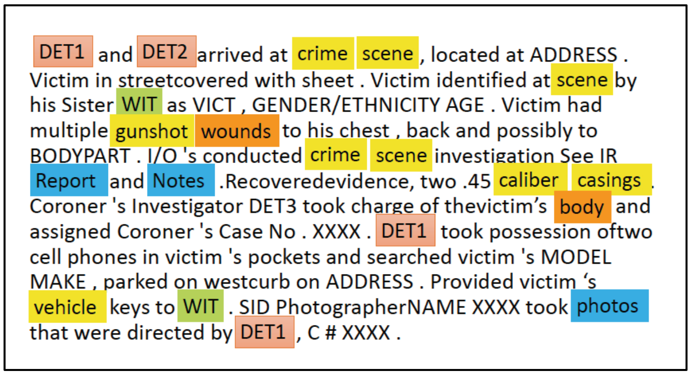
Appendix A: NLP Techniques that May Support Cold Case Data Analysis

The following section provides an overview of techniques that streamline the case review process for MDTs, providing more time for intensive investigation tasks.

Named-Entity Recognition (NER)

NER, a subset of information extraction, is a technique that identifies and categorizes entities. NER can be used for both structured text, such as digitized forms, or unstructured text, such as a witness narrative. NER can classify common terms that may be important in a variety of applications, such as names, locations, and times/dates; it can also identify domain-specific terms, such as “blood” and “semen,” as demonstrated in ***Exhibit 5***.

NER has multiple levels of complexity. In rules-based systems, the tools are “pre-trained” to recognize certain entities based on a rigid set of rules (e.g., extracting all numbers or text that appears to be in time/date form). This may be useful when the agency uses standard forms (like laboratory reports) or a common structure for narratives. However, if the criminal justice organization is trying to extract data from long, text-based datasets or if they want to extract words that are relevant for a cold case investigation, the organization will need to train tools enabled with NER using machine learning. An important first step of training this tool is identifying keywords (phrases) that a tool should look for (e.g., “DNA”).



DET1 and DET2 arrived at crime scene , located at ADDRESS . Victim in streetcovered with sheet . Victim identified at scene by his Sister WIT as VICT , GENDER/ETHNICITY AGE . Victim had multiple gunshot wounds to his chest , back and possibly to BODYPART . I/O 's conducted crime scene investigation See IR Report and Notes .Recoveredevidence, two .45 caliber casings . Coroner 's Investigator DET3 took charge of thevictim’s body and assigned Coroner 's Case No . XXXX . DET1 took possession oftwo cell phones in victim 's pockets and searched victim 's MODEL MAKE , parked on westcurb on ADDRESS . Provided victim ‘s vehicle keys to WIT . SID PhotographerNAME XXXX took photos that were directed by DET1 , C # XXXX .

Exhibit 5. Example output of named-entity extraction to extract relevant information from a case chronology.[7]

NER techniques can help agencies (1) automatically identify names, dates, and domain-specific terms (e.g., semen, blood) from digitized text files; (2) identify where these terms show up in the case files and help direct the user to this information, reducing time needed to manually sift through files; and (3) summarize high-level information from a case.

Relation Extraction

Relation extraction is an NLP technique that uses semantics to identify relationships between words in a sentence. This builds on NER to relate words to each other and understand their context within a sentence. For example, this technique can use context within a sentence to suggest that a person may have a particular role (e.g., suggest that Person A and Person B are witnesses to the crime). Relation extraction requires some pre-processing steps, such as processes that break down words into domains.

Relation extraction and topic modeling techniques can help agencies presumptively understand the roles and relationships between individuals mentioned in a case file such as person–person relationships, person–organization relationship, or organization–organization relationship, among others.[8] It can also help agencies understand case connections through link analysis, which establishes higher-order relationships among entities in a visual analysis tool. Link analysis, shown in ***Exhibit 6***, can draw interesting connections within and across different cases.

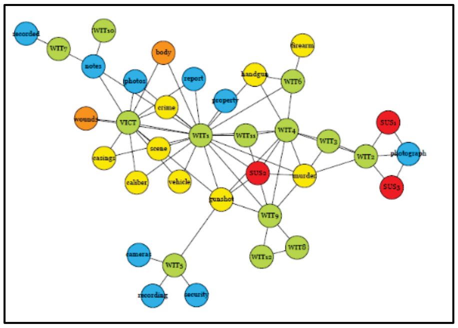


Exhibit 6. Example of a knowledge graph built from understanding relationships between different entities in a case chronology.[7]

Topic Modeling

Topic modeling scans a set of text files for similar word groups and expressions to cluster them based on shared characteristics. Tools using topic modeling employ pattern recognition (counting similar word patterns, like word frequency and distance between words) to identify where similarities exist between cases. This technique uses unsupervised learning, meaning that the algorithm is not "trained" on a set of terms to identify. Rather, the output is simply "what keywords or phrases are common in this dataset?"

Topic modeling techniques can help agencies identify similar details across cases that may not be readily apparent.

Text Classification

Text classification builds on topic modeling and sorts documents or text into categories outlined by the agency.[9] This technique can identify specific clusters of words or expressions but requires supervised algorithms, trained by a manually tagged dataset. The training helps the algorithm learn key words that indicate important information about a case.

Text classification techniques can classify cold cases (or documents within cold cases) with a set of tags that indicate relevant information about the case (e.g., whether it has a sexual assault nexus).

Appendix B: Additional AI-Based Techniques Adjacent to Text Analytics

NLP is used in many applications and workflows to help users find pertinent search results, use large amounts of data, and automate labor intensive processes. Many NLP-enabled tools can help agencies mine their unstructured text data for key insights; however, some tools have limited return on investment and may not be well-suited for cold case investigations. Some NLP applications include intelligent document processing and text summarization.

Intelligent Document Processing (IDP)

IDP is an automated workflow that digitizes and extracts specified data from a set of documents.[10] Like other digitization tools, IDP leverages optical character recognition to translate typed documents or forms into machine-readable text and then uses pattern extraction tools to extract important information from this data, such as names, dates, and locations. With the help of machine learning and training, IDP tools can extract information from a variety of document formats, but the technology is often used in industries that process semi-structured forms with few changes to the document structure over time. This technology has been used frequently in industries where near 100% accuracy is vital to the success of the tool, like banking and accounting, which requires a significant amount of training and human-in-the-loop testing. Consequently, these tools can take up to $500,000 to set up. Although IDP may be able to digitize and pull out some insights in the workflow, this approach is very expensive because it is used to attain a near-perfect accuracy score and is engineered for semi-structured or structured documents.

Text Summarization

Automatic text summarization tools compress a large amount of unstructured data into key insights that can be easily ingested. These tools take either an (1) extractive approach, which identifies significant sentences and aggregates them into a summary or (2) abstractive approach, where the tool interprets and summarizes the content. These tools can condense a large corpus of data into easy-to-read bullets or paragraphs, which may help investigators understand the content of a case file more quickly. There are many free or low-cost automatic summarizer tools online, such as QuillBot, Resoomer, and SMMRY, but most tools are not domain-specific, and these summarization tools are often imprecise or unable to extract key information. Query-based summarization is an active area of research that aims to extract important details out of scientific papers, or multiple documents. As this area develops and is ultimately transitioned into open-source and commercially available products, it may be valuable to the cold case application.

Appendix C: Glossary of Terms

Artificial Intelligence (AI): Leverages computers and machines to mimic both the problem-solving and decision-making capabilities of a human.[11]

Deep Learning: A subfield of machine learning; a neural network of three or more layers that simulates the behavior of a human brain to learn from large amounts of data.[12]

Intelligent Document Processing (IDP): Uses a combination of AI-powered automation and machine learning to classify unstructured documents, extract information, and validate the data.[13]

Link Analysis: Sometimes referred to as graph visualization or network visualization. Visually presents networks of connected entities, typically representing specific data points, as nodes and links.[14]

Machine Learning: A branch of AI and computer science focusing on data and algorithms that mimics the way humans learn. This allows the algorithm to gradually improve on its accuracy.[15]

Named-Entity Recognition (NER): An algorithm that scans through a dataset and defines important people, organizations, products, places, or general keywords. Once this information has been extracted, the algorithm helps with the automatic categorization.[16]

Natural Language Processing (NLP): A branch of computer science, specifically AI, that is concerned with giving computers the ability to understand and process text and spoken words in a similar manner that humans can.[17]

Neural Networks: A subset of machine learning and AI algorithms with a structure that mimics the human brain. Neural nets are composed of nodes, containing both an input and output, that are connected to each other.[18]

Relation Extraction: The task of extracting semantic relationships from text, usually occurring between two or more entities of a specific type, including people, places, and organizations.[19]

Semi-Supervised Tools/Learning: Branch of machine learning that attempts to solve problems that require both labeled and unlabeled data.[20]

Structured Data: Data that are highly organized and easily understood by machine learning. Most often categorized as quantitative data, these are the data with which we are most used to working. Examples include names, addresses, structured forms, or existing in predefined formats.[21, 22]

Supervised Tools/Learning: Machine-learning approach that uses labeled datasets. These datasets are trained to supervise algorithms into classifying data or predicting outcomes by using labeled inputs and outputs. Allows the model it is training to measure accuracy and continue to learn over time.[23]

Text Analytics: Automated process of translating unstructured text into quantitative data to uncover insights, trends, and patterns. Can be combined with data visualization tools to allow companies to better understand information and make informed decisions. Can be interchangeably used with text mining and text analysis.[24]

Text Classification: Also known as text tagging, it is the process of categorizing text into organized groups using NLP. This categorization can be based on content or from a set of predefined tags.[25]

Topic Modeling: Unsupervised machine-learning technique that is capable of scanning a set of documents, detecting both word and phrase patterns within the documents, and automatically clustering words and expression to best characterize the set of documents.[8]

Unstructured Data: Data that are difficult to deconstruct because they have no predefined model; therefore, these data cannot be organized into relational databases. Most often categorized as qualitative data. Examples include text, video files, data not contained in a structured form, and social media posts.[26]

Unsupervised Tools/Learning: Uses machine-learning algorithms to analyze and cluster unlabeled datasets. Typically, these algorithms will discover hidden patterns in data without the need for human intervention. They work on their own to discover the inherent structure of unlabeled data.[27]

References

1. Forensic Technology Center of Excellence. "In-brief report series: beyond DNA – sexual assault investigations." Last modified January. Accessed May 19, 2022. National Institute of Justice, https://forensiccoe.org/beyond-dna-reports-sexual-assault-reform/.
2. U.S. Department of Justice, Federal Bureau of Investigation. "FBI — Offenses Cleared." Accessed Mar 30, 2022. Federal Bureau of Investigation, https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/clearances.
3. Janani, R., and Vijayarani, S. *Text analytics in big data environments*. CRC Press, 2022.
4. Rigano, C., and National Institute of Justice. "Artificial Intelligence." Last modified October 8. Accessed February 25, 2022. https://nij.ojp.gov/topics/artificial-intelligence#recent-faqs-how-does-artificial-intelligence-learn.
5. Palantir Technologies, and National Center for Missing & Exploited Children. "Fighting child exploitation with big data." Accessed February 25, 2022. https://www.palantir.com/ncmec/.
6. Tauber, Alejandro. "How the Dutch police are using AI to unravel cold cases." Last modified May 23. Accessed February 25, 2022. The Next Web, https://thenextweb.com/news/how-the-dutch-police-is-using-ai-to-unravel-cold-cases.
7. Pandey, R. , Brantingham, P. J. , D., Uchida C., and Mohler, G. "Building knowledge graphs of homicide investigation chronologies." Paper presented at the 2020 International Conference on Data Mining Workshops (ICDMW), 2020.
8. Pascual, Federico. "Topic Modeling: an introduction." Last modified September 26. Accessed February 25, 2022. Monkey Learn, https://monkeylearn.com/blog/introduction-to-topic-modeling/.
9. Eggers, William D., Malik, Neha, and Gracie, Matt. "Using AI to unleash the power of unstructured government data: applications and examples of natural language processing (NLP) across government." (16 January 2019).
10. Team High Peak. "The Beginner's Guide To Intelligent Document Processing (IDP)." Last modified February 26. Accessed February 25, 2022. Medium.com, https://becominghuman.ai/the-beginners-guide-to-intelligent-document-processing-idp-5cc88b8de425.
11. IBM Cloud Education. "Artificial Intelligence (AI)." Last modified June 3. Accessed February 25, 2022. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.
12. IBM Cloud Education. "Deep Learning." Last modified May 1. Accessed February 25, 2022. IBM Cloud Education, https://www.ibm.com/cloud/learn/deep-learning.
13. IBM Cloud Education. "What is document processing?" Last modified October 26. Accessed February 25, IBM Cloud Education, https://www.ibm.com/cloud/blog/document-processing.
14. Cambridge Intelligence. "Link analysis: unlock the insight in your connected data with powerful link analysis." Accessed February 25, 2022. Cambridge Intelligence, https://cambridge-intelligence.com/why-link-analysis/.
15. IBM Cloud Education. "Machine Learning." Last modified July 15. Accessed February 25, 2022. IBM Cloud Education, https://www.ibm.com/cloud/learn/machine-learning.
16. GÜBÜR, Koray Tuğberk. "Named Entity Recognition: definition, examples, and guide." Last modified July 12. Accessed February 25, 2022. Holistic SEO, https://www.holisticseo.digital/theoretical-seo/named-entity-recognition/.
17. IBM Cloud Education. "Natural Language Processing (NLP)." Last modified July 2. Accessed February 25, 2022. IBM Cloud Education, https://www.ibm.com/cloud/learn/natural-language-processing.
18. IBM Cloud Education. "Neural Networks." Last modified August 17. Accessed February 25, 2022. IBM Cloud Education, https://www.ibm.com/cloud/learn/neural-networks.
19. Ruder, Sebastian. "Relationship Extraction." Accessed February 25, 2022. https://nlpprogress.com/english/relationship_extraction.html.
20. Meel, Vidushi. "What is semi-supervised Machine Learning? A gentle introduction." Accessed February 25, 2022. viso.ai, https://viso.ai/deep-learning/semi-supervised-machine-learning-models/.
21. Smallcombe, Mark. "Structured vs Unstructured Data: 5 key differences." Last modified January 3. Accessed February 25, 2022. Integrate.io, https://www.integrate.io/blog/structured-vs-unstructured-data-key-differences/.

NIJ Forensic Technology Center of Excellence
Visit us at
www.forensiccoe.org | ForensicCOE@rti.org | 866.252.8415
RTI International
3040 E. Cornwallis Road PO Box 12194, Research Triangle Park, NC 27709 USA

@ForensicCOE
#FTCoE

Sharing Knowledge | Advancing Technology | Addressing Challenges

22. Pickell, Devin. “Structured vs Unstructured Data – What’s the Difference?” Accessed February 25, 2022. G2, https://www.g2.com/articles/structured-vs-unstructured-data.

23. Delua, Julianna. “Supervised vs. unsupervised learning: what’s the difference?” Last modified March 12. Accessed February 25, 2022. IBM Cloud, https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

24. Pascual, Federico. “Text Analytics Basics: a beginner’s guide.” Last modified November 20. Accessed February 25, 2022. Monkey Learn, https://monkeylearn.com/blog/what-is-text-analytics/.

25. Monkey Learn. “What is text classification?” Accessed February 25, 2022. Monkey Learn, https://monkeylearn.com/what-is-text-classification/.

26. MonkeyLearn. “Text analytics basics: A beginner’s guide.” Accessed February 28, 2022. https://monkeylearn.com/blog/what-is-text-analytics/.

27. Pascual, F. “Topic modeling: An introduction.” Accessed Mar 30, 2022. MonkeyLearn, Inc., https://monkeylearn.com/blog/introduction-to-topic-modeling/.

Published November 2022

More Information

FTCoE Contact

Jeri Ropero-Miller, PhD, F-ABFT
Director, FTCoE, RTI International
jerimiller@rti.org

Jennifer Love, PhD, D-ABFA
Physical Scientist
Office of Investigative and Forensic Sciences
Jennifer.love@usdoj.gov

Technical Contacts

Patricia Melton, PhD
RTI International
pmelton@rti.org

Rebecca Shute, MS
RTI International
rshute@rti.org

Disclaimer

The NIJ FTCoE, led by RTI International, is supported through a Cooperative Agreement from the NIJ (2016-MU-BX-K110), Office of Justice Programs, U.S. Department of Justice. Neither the U.S. Department of Justice nor any of its components are responsible for, or necessarily endorse, this in-brief. NIJ is the research, development, and evaluation agency of the U.S. Department of Justice. NIJ is dedicated to improving knowledge and understanding of crime and justice issues through science. NIJ provides objective and independent knowledge and tools to inform the decision-making of the criminal and juvenile justice communities to reduce crime and advance justice, particularly at the state and local levels. The NIJ Office of Investigative and Forensic Sciences (OIFS) is the federal government’s lead agency for forensic science research and development. OIFS's mission is to improve the quality and practice of forensic science through innovative solutions that support research and development, testing and evaluation, technology, information exchange, and the development of training resources for the criminal justice community.

Public Domain Notice

Suggested Citation

Shute, Rebecca, Brailey Faris, Matthew Mecray, Mikalaa Martin, Ashley Rodriguez, Shannon Krauss, and Patricia Melton. “Digital Transformation in Cold Cases: The Application of Text Analytics. Forensic Technology Center of Excellence.” U.S. Department of Justice, National Institute of Justice, Office of Investigative and Forensic Sciences, November 2022.

NIJ | National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Forensic Technology
CENTER OF EXCELLENCE

RTI
INTERNATIONAL