Evidence Data Preprocessing for Forensic and Legal Analytics

Sundar Krishnan *skrishnanl@shsu.edu*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Narasimha Shashidhar *karpoor@shsu.edu*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Cihan Varol *cvarol@shsu.edu*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

ABM Rezbaul Islam *ari014@shsu.edu*
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Abstract

Electronic evidential data pertaining to a legal case, or a digital forensic investigation can be enormous given the extensive electronic data generation mechanisms of companies and users coupled with cheap storage alternatives. Working with such volumes of data can be tasking, sometimes requiring matured analytical processes and a degree of automation. Once electronic data is collected post eDiscovery hold or post forensic acquisition, it can be framed into datasets for analytical research. This paper focuses on data preprocessing of such evidentiary datasets outlining best practices and potential pitfalls prior to undertaking analytical experiments.

Keywords: eDiscovery, Electronic Stored Information, Digital Evidence, Digital Forensics, Digital Forensic Analytics, Legal Analytics, Machine Learning, Preprocessing, Natural Language Processing.

1. INTRODUCTION

Computers, mobile devices, smartphones, medical devices, the Internet Of Things and other electronic devices maybe used for committing crime, making law enforcement to leverage digital forensics to fight crime. These devices store, receive and transmit data that can be of critical evidential value to an investigation or legal arguments [1]. Digital evidence is now used to prosecute both civil and criminal cases when the evidence pile involves electronic devices. Ultimately, digital evidence for a case should be admissible in court and its significance explained to a jury. Digital forensic experts assist counsel through legally reliable methods to ensure Digital Evidence's admissibility in both civil and criminal cases. Unfortunately, the volume of digital evidence can be overwhelming for digital forensic experts given the growth of technology [2]. A potential solution can be the use of analytics such as machine learning and artificial intelligence to assist in the review of evidence. After the forensic acquisition of evidence from an electronic device, digital forensic experts can export a read-only copy of raw evidence data from forensic tools to conduct analytical experiments.

In the legal system, discovery is the process that governs the right to obtain and the obligation to produce non-privileged matter relevant to any party's claims or defenses. eDiscovery is the discovery process applied to Electronically Stored Information (ESI) or case data, such as emails, digital data from the Internet, computer files, databases, etc. The growing emphasis on paperless files and collaborative computer systems coupled with connectivity to the Internet and cheap Cloud storage has created even greater volumes of electronic information. A legal case ESI can be voluminous and can be challenging especially during the review stages [3]. This means that attorneys and other legal professionals will have to deploy and learn new technology to quicken the pace of eDiscovery while maintaining the quality of work. Legal professionals can use the Electronic Discovery Reference Model (EDRM) [4] as a starting point and is widely considered as the definitive framework for the eDiscovery process.

Many eDiscovery solutions/tools have focused on improving collection efficiency and reducing data review effort for long. Legal analytics is the management process of extracting actionable knowledge from data to assist in-house legal leaders and decision-makers [5]. Few use cases for legal analytics involve eDiscovery efficiency, motion forecasting, process improvement, legal strategy, comparative legal costs, billing optimization, settlement award, resource management, and financial operations. To summarize, legal analytics tools help lawyers make data-driven decisions on which to build their legal strategies [6]. During eDiscovery, once relevant case data is collected and loaded into a storage platform, legal teams begin reviewing the data. Data reviews can be time-consuming and are the bulk of litigation costs. Thus, leveraging technology through an established framework can greatly help with speed and accuracy during reviews. In 2012, EDRM proposed Technology-Assisted Review (TAR) [7] and has since been steadily gaining popularity with the industry as an essential tool in eDiscovery. The TAR framework (also known as predictive coding) refers to a document review approach in eDiscovery that leverages computer algorithms to identify and tag potential documents based on keywords and metadata. TAR in it's original form is a multi-step process spanning anywhere between 6 to 10 steps depending on whether Simple Active Learning (SAL) or Simple Passive Learning (SPL) was being used. The second generation of Technology-Assisted Review (TAR 2.0), however, consists of Continuous Active Learning (CAL), which enables a system to continuously analyze the machine learning results (in the background) as humans review documents without the need to begin by analyzing static, randomized samples [8]. The result is a non-iterative and continuously improving implementation of TAR as the review progresses by re-ranking the entire data set with each new batch of data [8]. Continuous Active Learning indicates that the system uses the updated model to continuously promote case documents to the top of the review queue that has the highest probability of being responsive to the case [9]. Thus, TAR 2.0 has many advantages over TAR 1.0. In TAR 1.0, experts do the initial training, and it is less effective because it cannot learn from subsequent decisions. TAR 1.0 also cannot handle rolling productions without having to start over [10]. In TAR 2.0, all human review decisions automatically train and update the system predictions as new human classifications are made.

Technology Assisted Review (TAR) has established itself into standard e-discovery practices with a key benefit of expediting the document review process. TAR has also garnered favor with judges familiar with its benefits [11] and also has judges refusing to compel parties to apply TAR [12]. Analytical techniques such as Machine Learning (ML) (supervised or un-supervised), Artificial Intelligence (AI), Deep Learning, Neural Networks, Statistical approaches, etc. fall under TAR technology umbrella. Since recently, these techniques have gained popularity with legal firms and eDiscovery solutions vendors with a goal to expedite the organization and prioritization of document collection and minimize review efforts. These techniques help save costs and reduce time in helping to identify relevant data. ML-based solutions such as Brainspace [13] can perform conceptual clustering by reading case documents, searching for relevant words, and clustering them into groups based on their contents [14]. AI is a very useful assistant when helping to identify relevant data by leveraging supervised learning. However, these techniques have limitations as they clearly do not run the investigation but, merely assist in speeding up the overall process.

Data analytics is a broad term that refers to the use of various techniques that find meaningful patterns, predict the future, and give insights into data. Data analytics is not new to digital forensics or to the legal world and can be as simple as employing statistics in decision making. Few enabling fields of data analytics are data science and data engineering. Data science is a process of testing, evaluating, and experimenting to create and apply new data analytic techniques. Data engineering makes data useful by helping structure data making it easier for application and human consumption. Data analytics has greatly manifested in the last few years as we focus more on business intelligence and the real-time analysis of data [15]. The explosion of smartphone usage, coupled with easier Internet connectivity and low costs of Cloud storage, has converted data analytics into a buzzword. The need to derive meaningful insights into customer or business data has pushed disciplines such as text/image mining, predictive modeling, etc. The technical aspects of analytics can be found in the emerging fields of machine learning techniques such as neural networks, decision trees, logistic regression, linear/multiple regression analysis and classification. All of these disciplines require clean raw data for input and the process of cleaning/transforming raw data is known as preprocessing. Often raw data is likely to be imperfect, noisy, inconsistent, and sometimes redundant, making it unfit for analysis. Analytical experiments greatly depend on the quality of input data, and as such, results can be skewed or incorrect if data was not preprocessed correctly prior to applying algorithms. These days, law firms, eDiscovery vendors, legal and forensic researchers, have all started to venture into experimenting with advanced analytical techniques such as Machine Learning and Artificial Intelligence. Legal and forensic analytical experiments can be around actual digital forensic investigations of a case, reviews of case ESI during eDiscovery, staged experiments for research and process optimization. There have been promising results when applying these advanced analytical techniques in legal eDiscovery yielding in direct financial advantages. However, there exists caution in the legal industry and digital forensics investigations when leveraging such techniques as applying analytics is still in a nascent stage with courts being the ultimate proving ground in validating their use. In this article, the authors share their best-practices when preparing for such advanced analytical experiments in a legal setting under TAR or within a digital forensic investigation scope. Suggested best practices are guidelines that can lower risk and improve the statistical model's efficiency and accuracy when employing analytical techniques such as Machine Learning or Artificial Intelligence to work in a legal setting or forensic investigation.

2. RELATED WORK

Data preprocessing and dimensionality reduction is an integral step in any analytical experiment leveraging statistics, machine learning, artificial intelligence, neural networks, etc. The number of features, quality of input data, and the useful information that can be derived from it directly impacts the ability of the algorithms and eventually the result. A typical use case in analytic experiments in eDiscovery is around reviewing emails within the case ESI. Email preprocessing can help identify spam, categorize emails and mitigate phishing attacks. Ruskanda [16] studied the effect of preprocessing of emails on spam email detection techniques using supervised spam classifier algorithms: Na¨ıve Bayes and Support Vector Machine. Kumara et al. [17] propose an enhanced data preprocessing approach for multi-category email classification by ignoring the signatures on emails, special characters, and unwanted words. Their proposed model was evaluated using various classifiers and showed that the proposed data preprocessing to email classification is superior to the existing approach. Emails can be complex to parse and process due to branching, forwarding, attachments, multiple languages, signatures, footers, disclaimers, auto-generated phishing warnings, URLs, etc. Tang et al. [18] in a cascaded approach, propose leveraging Support Vector Machines (SVM) to clean up emails by addressing non-text filtering, paragraph normalization, sentence normalization, and word normalization. Emails branch and can sometimes render a partial picture of the whole conversation. A single longest thread alone can only track a linear, back-to-back conversation. Instead, to factor the whole conversation, branching emails should be considered and grouped [19]. Another focus area of legal analytics during reviews is data from social media that can be littered with words from multiple languages, jargon, code words, abbreviations, shortened words, etc. Uysal et al. [20] examine the impact of preprocessing on text classification using benchmark datasets. They concluded that the choice

and combinations of preprocessing tasks may provide a significant improvement on classification accuracy depending on the domain and language. Kantepe et al. [21] propose a preprocessing framework for Twitter bot detection with reasonable accuracy using a machine learning supervised classification approach. Etaiwi et al. [22] investigate the effects of preprocessing steps on the accuracy of reviews spam detection by applying machine-learning algorithms against a labeled dataset of hotel reviews. Data from social media can be complex simply due to multiple languages used, sharing, liking, commenting, etc. There exists a gap in literature focusing on preprocessing challenges and best-practices when working on analytical experiments or research with forensic evidence and legal case data. While existing literature focusses on generalized approaches towards various data preprocessing techniques, algorithms, etc. there is little contribution towards applying such methodology towards industry specific use-cases. In this paper, the authors discuss best-practices and potential issues for legal and forensic data analysts during data preprocessing when working in forensic and legal investigations or analytical tasks.

3. DATA PREPROCESSING FOR FORENSIC AND LEGAL ANALYTICS

A caseload of digital evidence can be viewed as a data-lake that can translate into meaningful datasets for analytical experiments. To understand the depth of analytical algorithms, the features (attributes or variables) in the evidence/case data, and what they represent are to be well understood. This section delves into best practices when preparing for analytical experiments using evidentiary case data during legal analytics or forensic investigations.

3.1 Research Methodology

The methodology of this paper includes reviewing existing literature, examining best-practices and potential pitfalls during data preprocessing in forensic and legal investigations in addition to following current industry trends.

3.2 Identify Analytical Aim/Problem/Objective

Like any analytical experiments, legal and forensic analytics will need to identify aims to accomplish or problems to be solved prior to the start of experiments. They can help devise a strategy and identify the data that needs to be collected. Aims or problems are usually derived from the investigation scope, forensic protocol, or legal case scope. In a legal case, scope can be defined as the extent of ESI discovery that the parties agree to produce for the case and is generally defined by the Federal Rule of Civil Procedure 26(b)(1) [23]. During a digital forensic investigation, the scope and forensic protocol can be obtained from the investigation plan, security incident response or warrants. Scope limitations may be in effect due to time availability, forensic skills availability, forensic tool availability, budget, privacy or opposing interests. Figure 1 highlights the sources for deriving Aim/Problem/Objective in legal and forensic analytics.

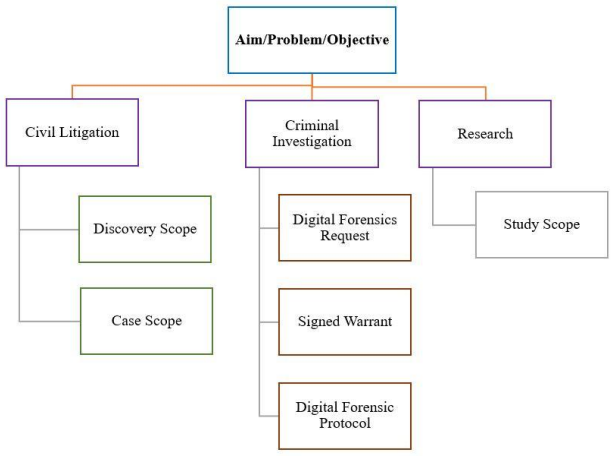

FIGURE 1: Sources of analytical aim/problem/objective in legal and forensic analytics.

3.3 Understanding Case or Evidence Data

To gain actionable insights into a legal case or forensic investigation, the appropriate data from case ESI or evidence must be sourced and cleansed prior to conducting analytical experiments. Care must be taken not to spoil the data by hampering its integrity, and, thus a true, verifiable copy of the data may be used for analytical experiments. There are two key stages of data Understanding: Assessment and Exploration. The first step is assessment during which, availability, format, storage, source, features, relevance, quality, reliability, etc., are explored. During the exploration step, missing values, outliers, bias, balance, etc., are explored. Case ESI data or evidence data post forensic acquisition can arrive from various devices/sources and in different raw formats. Data can be uploaded into a database or into spreadsheets for easy exploration. Statistical formulae can be used to further explore balance, mean, variance, etc. Feature engineering can then help normalize and scale data.

Few types of analytics that are having a significant impact on eDiscovery and forensic investigations are Machine Learning (ML), Convolutional Neural Networks (CNN), and Natural Language Processing (NLP). Machine learning uses mathematical models to assess enormous datasets, make predictions and learn from feedback. NLP allows machines to "understand" natural human language, thereby enabling computers to effectively communicate in the same language as their users. Although NLP and its sister study, Natural Language Understanding (NLU) are constantly advancing in their ability to compute words and text, human language can be complex, ever-evolving, fluid, and inconsistent thereby presenting serious challenges that NLP is yet to completely overcome. Since case data can mostly comprise of text, NLP is a suitable technique that is commonly used. Table I outlines few challenges when working with text-based case data. Figure 2 shows potential issues with raw data of a legal case ESI.

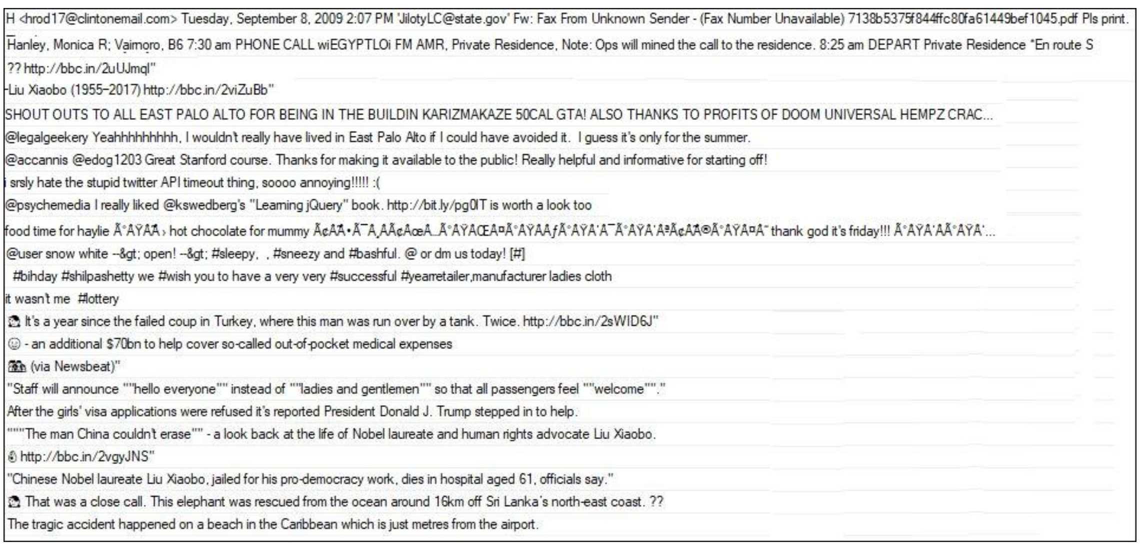
H <hrod17@clintonemail.com> Tuesday, September 8, 2009 2:07 PM 'JilotyLC@state.gov' Fw: Fax From Unknown Sender - (Fax Number Unavailable) 7138b5375f844ffc80fa61449bef1045.pdf Pls print.
Hanley, Monica R; Vaimoro, B6 7:30 am PHONE CALL wiEGYPTLOi FM AMR, Private Residence, Note: Ops will mined the call to the residence. 8:25 am DEPART Private Residence "En route S
?? http://bbc.in/2uUJmql"
Liu Xiaobo (1955–2017) http://bbc.in/2viZuBb"
SHOUT OUTS TO ALL EAST PALO ALTO FOR BEING IN THE BUILDIN KARIZMAKAZE 50CAL GTA! ALSO THANKS TO PROFITS OF DOOM UNIVERSAL HEMPZ CRAC...
@legalgeekery Yeahhhhhhhhh, I wouldn't really have lived in East Palo Alto if I could have avoided it. I guess it's only for the summer.
@accannis @edog1203 Great Stanford course. Thanks for making it available to the public! Really helpful and informative for starting off!
i srsly hate the stupid twitter API timeout thing, soooo annoying!!!!! :(
@psychemedia I really liked @kswedberg's "Learning jQuery" book. http://bit.ly/pg0IT is worth a look too
food time for haylie Ã°ÂŸÂÃ› hot chocolate for mummy Ã¢ÂÃ•Ã¯Â¸ÂÃ¢ÂœÂ…Ã°ÂŸÂŒÂ¤Ã°ÂŸÂÂƒÃ°ÂŸÂ'Â¯Ã°ÂŸÂ'ÂªÃ¢ÂÃ®Ã°ÂŸÂ¤Â¨ thank god it's friday!!! Ã°ÂŸÂ'ÂÃ°ÂŸÂ'...
@user snow white --> open! --> #sleepy, , #sneezy and #bashful. @ or dm us today! [#]
#bihday #shilpashetty we #wish you to have a very very #successful #yearretailer,manufacturer ladies cloth
it wasn't me #lottery
It's a year since the failed coup in Turkey, where this man was run over by a tank. Twice. http://bbc.in/2sWID6J"
☺ - an additional $70bn to help cover so-called out-of-pocket medical expenses
(via Newsbeat)"
"Staff will announce ""hello everyone"" instead of ""ladies and gentlemen"" so that all passengers feel ""welcome""."
After the girls' visa applications were refused it's reported President Donald J. Trump stepped in to help.
"""The man China couldn't erase"" - a look back at the life of Nobel laureate and human rights advocate Liu Xiaobo.
http://bbc.in/2vgyJNS"
"Chinese Nobel laureate Liu Xiaobo, jailed for his pro-democracy work, dies in hospital aged 61, officials say."
That was a close call. This elephant was rescued from the ocean around 16km off Sri Lanka's north-east coast. ??
The tragic accident happened on a beach in the Caribbean which is just metres from the airport.

FIGURE 2: Sample raw text in a case ESI or digital forensic evidence prior to preprocessing. Contains garbled characters, Unicode, email addresses, shorthand, slang, URLs, emoji and hashtags.

The use of programming languages, software, and automation technology can sometimes impact data integrity. Storage of raw case/evidence data on databases should be undertaken with caution to support Unicode, logos, signatures, image & video pixel resolution, gifs, VR media, etc. Database or file-system transactions should not alter the state of raw evidence data. For example, for processing Facebook data in Arabic or French language containing emoji (a true-copy from a case ESI or digital evidence) stored on a SQL Server database instance, should consider the schema (column-level) design for Unicode and multilingual language support. Similarly, transacting with this database using Python programming language to perform analytical research should be undertaken with caution as read writes into the database can accidentally ignore/suppress Unicode support, thereby impacting data integrity and experiment

results. Thus, a cursory glance at raw data should be undertaken before identifying and designing technology platforms for analytics.

3.3 Technology Selection

Digital Forensic tools, email processing tools, social media crawlers, eDiscovery solutions, and various other extraction/parsing tools are some of the technology-driven tools that can help extract and export data from case evidence. Not all tools export extracted data in the same format. Thus, for analytical experiments, data has to be collated into a single dataset with necessary features. Appropriate computer programs can be leveraged to legally obtain social media website data via their defined application programming interfaces (API). Relational databases can be used to collect and store data following which queries may be used to create datasets. Randomly, exported data from the tool will need to be validated against reported/observed evidence (device) data for tool accuracy and dependability. The assistance of data scientists, data engineers, statisticians, domain experts and Information Technology staff may be required when conducting any legal analytical experiments.

3.4 Digital Forensics

There exists an interplay between eDiscovery and digital forensics [24] when data from evidence will need to be forensically extracted for legal arguments and investigation. The collection phase of eDiscovery is when digital forensic professionals are often engaged to protect data integrity and to bring forth the data stored on digital evidence. Digital forensic tools export evidence data into various formats. Note that not all forensically acquired data (evidence) may be directly ready for analytical experiments. Images, audio, and video files may contain hidden data or be deep-fake needing to be suitably addressed. Few variations of legal analytical research may involve forensic investigations. For example, predicting friends using social media data or clustering documents related to a crime. During such research, the investigative skills of digital forensic professionals may be leveraged to validate results.

3.5 Identify Key Features

In a legal case-load of evidence, data within the evidence device/source is not always ready for immediate analytical experiments. Case evidence data often can be found as digital files from various software programs or plainly skimmed off the Internet. This makes identification of data within such data a prerequisite, as data can be generally voluminous and uncured. Key features (attributes or variables) of data will need to be identified for the legal case. Identifying key features ahead of an analytical experiment requires planning and assistance from technical experts on the case. Key features may start from a wish-list but should be scoped to translate into being technically feasible collection while mainlining data integrity all through the process. For example, if the case arguments hinge upon presence of the client at specific locations over a time, then details such as timestamps and geographical location from data are key features that need to be collected into datasets. In another example, if the case arguments hinge upon the use of a computer for certain Internet activities, features from case-data such as login data (of both computer and online websites such as timelines, authentication tokens, the identity used), web activity (timelines, posts, likes, dislikes, and comments) and geographical location data from network traffic may be of use. Ancillary features such as online responses from friends/strangers of the defendant/client may add noise and degrade the analytical algorithms in the experiments. Multiple datasets of such key features can be then prepared for individual analytical experiments.

3.6 Data Threads

Disentangling conversations mixed into a single stream of messages can create challenges unless properly handled and carved into detached yet linked data. Further complications arise when conversations are peppered with slang, abbreviations, URLs, etc. A common occurrence of such conversations are long email threads that are often the first to be reviewed during eDiscovery following “The Longest Thread Policy” [19]. An email thread is a group of emails all originating from the same email that branch off in many directions as receivers (copied or blind-copied) forward the email to different recipients. Sometimes, other email threads can interweave into threads that can complicate a walk. Slicing emails from threads for analytical experiments

can cause data loss or introduce noise. In some instances, senders may manually remove or edit certain email body when forwarding or replying. Such data loss should be monitored. Automation tools that help parse emails should be carefully chosen to report any such discrepancies. Similarly, conversations on social media platforms can branch (like a tree) into multiple senders and receivers. A conversation path must be identified to isolate actors/subjects, timelines, and their conversations. Improper handling of such lengthy strings of data can also lead to missing out on the context of the whole conversation. Parsing attachments, embedded videos or images in such threads can add to the complexity, thus requiring design considerations on datasets.

Description	Expression
Loan-words in English of foreign origin	bona fide ad nauseam, en masse, faux pas, fait accompli, modus operandi, persona non grata, quid pro quo bon voyage, pro bono, status quo, avatar, guru, chilly (means peppers in Indian language), hullabaloo, mulligatawny, Chop chop, Feng shui, Coolie, Nankeen (durable cloth in Mandarin)
Sarcasm	“Is it time for your medication or mine?” “My favorite thing to do at 5AM is to go to the Airport. How about you?” “That’s just what I needed today!”
Irony	“The fire station burned down” “The traffic cop got his license suspended because of unpaid parking tickets”
Errors in text or speech (Psycholinguistic classification like deletion, blends, addition, omission, etc. [25])	“Bake my bike” “He pulled a pantrum” “Both sicks are kids”
Colloquialisms and slang	“I’m fixin’ to go to the park” “Blimey” - exclamation of surprise, “Chockablock” - something that is completely filled, “Dodgy” - something less than safe or secure, “Lemon” - a purchase that is unreliable

TABLE 1: Common language and text limitations in case evidence data.

3.7 Data Correlation

Finding correlations in data from multiple data sources may be needed as part of analytical experiments. Correlation is like finding a pattern on wallpaper and is a statistical-based information analysis technique of analyzing relationships between two or more features (variables). For example, correlating data from sources such as company email and Facebook activity may be needed for legal arguments. In such situations, data for emails may be extracted from an exchange server or Microsoft 365 and Facebook data may be extracted from a smartphone. Creating datasets using both sources of data will need design insights and adequate planning.

3.8 Goodness of Fit

Model fitting is a measure of how well a machine learning model generalizes data that is similar to which it was trained for [26]. A good model fit is a statistical hypothesis test that of a model that accurately approximates the output when it is provided with unseen inputs. The goodness of fit of a statistical model describes how well it fits a set of observations. Over fitting a model captures the noise and outliers in the data along with the underlying pattern. Such models usually have high variance and low bias. Under fitting a model occurs when the model is unable to capture the underlying pattern of the data and is too simple. Such models usually have a low variance and a high bias. Bias and variance are key risks in analytical experiments and can be best addressed by implementing statistical best practices. Bias exists in all data driven experiments, but the question is how to identify and remove it from the experiment. Bias can skew results and might

negatively impact the effectiveness of the experiment's algorithms. To avoid bias, careful planning of the experiment is needed, and a balance between transparency and performance has to be maintained. Bias in analytical experiments can eventually derail a legal case.

3.9 Data Loss

Inadvertent data conversions can lead to data loss. Care should be taken in instances when emoji, glyphs, Unicode scalars, favicons, emoticons, nicknames, slang words, abbreviations, Anglicized language, etc. are embedded in text. Encoded conversations, embedded images or videos can change the meaning to a plain text conversation but may also hold a secret meaning for the intended targets. Data transformation, filtering, encoding, removing email appends (logos, banners, system-generated phishing warnings, printer ink-friendly messages), etc. can all lead to data loss. However, this must be documented and not adversely impact the aim of the analytical experiment.

3.10 Data Leakage

Often encountered during predictive analytics, data leakage is when information from outside the training dataset is used to create a model. This can be accidental sharing of information between the test and training data during the experiment, or during data preprocessing. Data Leakage can lead to false assumptions about the performance of the analytical model. Generally, if the analytical model is too good to be true, we should be suspicious.

3.11 Sensitive Data and Privacy

Sensitive data is any data such as personally identifiable information (PII), Protected Health Information (PHI), Payment Card Industry (PCI) data, Intellectual Property (IP), and other important business data. Analytical experiments may need to use such sensitive data. Legal firms have to comply with common industry regulatory standards for data protection and privacy such as; the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), Health Insurance Portability and Accountability Act (HIPAA), the Payment Card Industry Data Security Standard (PCI DSS), standards from the International Organization for Standardization (ISO), and others. Prior identification of sensitive data by manual or by leveraging pre-tuned industry tools is recommended. Specific use approvals from data custodians or identified authority is recommended prior to starting on analytic experiments. Processing of sensitive data through encryption, tokenization, redaction, masking, or de-identification maybe needed. For example, masking of last names may be required, or certain geographical location data may need to be obfuscated to protect privacy and identity. If so, these features may need to be dropped or encoded accordingly during analytical preprocessing. If authorized to use raw data for experiments, care must be taken for storage and distribution of results lest they accidentally expose sensitive data.

4. DATA MANAGEMENT DURING ANALYTICS

A disciplinary approach should be maintained during preprocessing and filtering of data when building a dataset. Multiple copies of data or datasets stored indiscriminately on storage drives/network can increase security and privacy risks. Industry best practices should be implemented, or organization policies followed when creating copies of case data. To avoid spoliation and accidental evidence corruption, a read-only copy of original raw evidentiary data should be carefully generated prior to use in any research or experiments.

4.1 Data Integrity

Data preprocessing steps can be lengthy when arriving at the best set of features for the analytical experiments. Care should be taken on data integrity as indiscriminate processing can truncate or manipulate data. For example, careless rounding of a float datatype or encoding a string datatype into a numeric datatype can impact the performance of the model and impact experiment conclusions. When exporting data off automation or forensic tools, similar caution should be employed lest the tools accidentally convert, format, or truncate data (data types). For example, when exporting timelines from a smartphone post digital forensic investigation, care

should be taken to maintain the date and time format of data, timezones especially when the device was used across countries. Transposing such data to adjust for the analytical experiment's needs should be undertaken with caution and documented.

4.2 Security and Access

Proper access (authorization and authentication) to data should be considered before the start of any analytical experiments. Access to data can be limited to read-only. Data shares with other teams should be part of authorization protocols. Similarly, reports and analysis from analytical experiments should be carefully shared with those who are authorized to receive them. Once analytical experiments are completed, authorization should be revoked to case data. Unless allowed by enterprise policy, use caution when sharing case data or analytical experiment results over emails or through enterprise messaging/chat applications. Industry best practices around security and privacy should be followed such as, implementing Data Loss Prevention (DLP) controls on endpoints and monitoring of network traffic.

4.3 Policy and Guidelines

Legal firms, eDiscovery/forensic practitioners, forensic labs, and vendors should ensure data management and governance, privacy, ethics and security policies are in place when working with case data. A separate policy and set of standards may be envisioned to address analytical research.

4.4 Backup and Retention

Plans for analytical research and experiments should follow enterprise backup and retention procedures. Pre-determined backup (storage) locations must be identified, and retention period defined.

4.5 Destruction

Upon completion or termination of analytical research and/or experiment(s) using case data, the concerned Information Technology or Security teams should be notified. Industry best practices, standards [27], [28] or enterprise defined policies may be employed for data clean-up destruction) processes to counter residual data. For example, if a Cloud based storage location or a portable storage-media were used as part of the analytical research and/or experiment(s), proper procedures must be followed to wipe the storage media or engage with the Cloud Service Provider to undertake the same. Likewise, systems used during the analytical research and/or experiment(s) should be subject to safe wiping policies and procedures.

5. CONCLUSION

Advanced analytical research and experiments are these days undertaken in-house by teams of data scientists with a background in legal, eDiscovery, Information Technology and Statistics. Forensic and legal analytics has come to the forefront of investigations and technology-assisted reviews given the recent focus in Machine Learning, Artificial Intelligence, and Deep Learning. In a legal case, digital evidence may be present as digital devices or Internet data. Extracting data off such evidence can be voluminous and can burden the review process during eDiscovery. Advanced analytical processing by digital forensic and legal professionals can come to the rescue of winnowing and interpreting large volumes of evidence data for establishing patterns, intent, and motives. Also, forensic, and legal analytical approaches can be used in forensic investigations to reduce evidence search time, gain insight into suspect's activities, clustering suspect profiles, optimize legal costs, case billing, motion prediction, legal strategizing, etc. All legal analytical research or experiments require data as inputs and raw data may not always be of the best quality for direct consumption. This paper outlines best practices and approach for preprocessing legal data prior to forensic and legal analytics. Leveraging analytics can greatly assist in manual case reviews and investigations but should not be considered as their replacement and solely relied upon as applying analytics is still considered as nascent in legal minds. It can be safely predicted that forensic and eDiscovery experts will soon need to add analytical and statistical skills to their knowledgebase to leverage them in their work and explain

the significance of these fields to a jury when offering expert opinions and interpreting investigation findings. In future work, the authors propose to focus on assessing the performance of legal analytical techniques to test and confirm the accuracy of preprocessing of evidentiary case data.

6. REFERENCES

[1] "Digital Evidence and Forensics." Internet: https://nij:ojp:gov/digital-evidence-and-forensics, [Mar. 02, 2021].

[2] D. Quick and K. K. R. Choo, Dec 2014, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," Digit. Investig., [On-line] vol. 11, no. 4, pp. 273–294,Available: https://www.sciencedirect.com/science/article/abs/pii/S1742287614001066, [Mar. 02, 2021].

[3] S. Krishnan, A. Neyaz, and N. Shashidhar, 2019, "A Survey of Security and Forensic Features In Popular eDiscovery Software Suites,". [On-line]. Available: https://www:cscjournals:org/manuscript/Journals/IJS/Volume10/Issue2/IJS-152:pdf, [Mar. 02, 2021].

[4] Electronic Discovery Reference Model, Internet: https://edrm.net/resources/frameworks-and-standards/, [Mar. 02, 2021].

[5] "Legal Analytics.", Internet: http://www:argopoint:com/legalanalytics, [Mar. 02, 2021].

[6] "What is Legal Analytics?", Internet: https://www:lexisnexis:com/community/lexis-legal-advantage/b/insights/posts/what-is-legal-analytics, 2019, [Mar. 02, 2021].

[7] EDRM, "Technology Assisted Review.", Internet: https://edrm:net/resources/frameworks-and-standards/technologyassisted-review/, [Mar. 04, 2021].

[8] S. Kernisan, "TAR 1.0 or TAR 2.0: Which method is best for you?", Internet: https://www:casepoint:com/blog/tar-1-0-versus-tar-2-0/, [Mar. 04, 2021].

[9] G. Taranto, "The Evolution of TAR", Internet: https://www:law:com/2020/12/31/the-evolution-of-tar/?slreturn=20210110063112, 2020, [Mar. 04, 2021].

[10] J. Kerry-Tyerman, "Why Machine Learning Matters in Ediscovery", Internet: https://www:everlaw:com/blog/2018/01/03/machine-learning-in-ediscovery/, 2018, [Mar. 04, 2021].

[11] Casetext, "Moore v. Groupe, 868 F. Supp. 2d 137", Available: https://casetext:com/case/moore-v-groupe, 2012, [Mar. 04, 2021].

[12] Hyles v. City of New York et al, No.1:2010cv03119 - Document 97 (S.D.N.Y. 2016), Available:https://law:justia:com/cases/federal/district-courts/new-york/nysdce/1:2010cv03119/361399/97/, 2016, [Mar. 06, 2021].

[13] Brainspace: Make Smarter, Faster, & More Informed Decisions., Available: https://www:brainspace:com/, [Mar. 07, 2021].

[14] Artificial intelligence and machine learning in e-discovery and beyond., Available: https://www2:deloitte:com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery:html, [Mar. 07, 2021].

[15] L. Wilson, “Enterprise AI: Data Analytics, Data Science and Machine Learning”, Available: https://www:cio:com/article/3342421/enterprise-ai-data-analyticsdata-science-and-machine-learning:html, 2018, [Mar. 07, 2021].

[16] F. Z. Ruskanda, Mar 2019, “Study on the Effect of Preprocessing Methods for Spam Email Detection,” Indones. J. Comput., [On-line] vol. 4, no. 1, p. 109, Available: http://www:mail-abuse:com/, [Mar. 07, 2021].

[17] B. A. Kumara, M. M. Kodabagi, T. Choudhury, and J.-S. Um, Jan 2021, “Improved email classification through enhanced data preprocessing approach,” Spat. Inf. Res., [On-line] pp. 1–9, Available: https://link:springer:com/article/10:1007/s41324-020-00378-y, [Mar. 07, 2021].

[18] J. Tang, H. Li, Y. Cao, and Z. Tang, 2005, “Email data cleaning,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. New York, New York, USA: ACM Press, [On-line] pp. 489–498, Available: http://portal:acm:org/citation:cfm?doid=1081870:1081926, [Mar. 07, 2021].

[19] J. Greer, “Email Threading in eDiscovery: The Longest Thread Policy,” Internet: https://www:digitalwarroom:com/blog/emailthreading-ediscovery-problems-with-longest-thread, 2019, [Mar. 08, 2021].

[20] A. K. Uysal and S. Gunal, Jan 2014, “The impact of preprocessing on text classification,” Inf. Process. Management., [On-line] vol. 50, no. 1, pp. 104–112, Available: https://doi.org/10.1016/j.ipm.2013.08.006., [Mar. 08, 2021].

[21] M. Kantepe and M. C. Ga˜niz, Oct 2017, “Preprocessing framework for Twitter bot detection” in 2nd Int. Conf. Comput. Sci. Eng. UBMK 2017. Institute of Electrical and Electronics Engineers Inc., [On-line] pp. 630–634, Available: https://doi.org/10.1109/UBMK.2017.8093483, [Mar. 12, 2021].

[22] W. Etaiwi and G. Naymat, Jan 2017, “The Impact of applying Different Preprocessing Steps on Review Spam Detection,” in Procedia Computer Science., vol. 113. Elsevier B.V., [On-line] pp. 273–279., Available: https://doi.org/10.1016/j.procs.2017.08.368., [Mar. 12, 2021].

[23] Rule 26. Duty to Disclose; General Provisions Governing Discovery— Federal Rules of Civil Procedure — US Law — LII / Legal Information Institute., Internet: https://www:law:cornell:edu/rules/frcp/rule 26#rule 26 a 1 B, [Mar. 10, 2021].

[24] S. Krishnan and N. Shashidhar, Mar 2021, “Interplay of Digital Forensics in eDiscovery,” IJCSS, [On-line] vol. 15, issue 2, pp 19-44, Available: https://www.cscjournals.org/manuscript/Journals/IJCSS/Volume15/Issue2/IJCSS-1602.pdf, [Mar. 19, 2021].

[25] “Speech error - Wikipedia.”, Internet: https://en:wikipedia:org/wiki/Speech error, [Mar. 22, 2021].

[26] Definition: Model fitting, Internet: https://www:educative:io/edpresso/definition-model-fitting, [Mar. 25, 2021].

[27] NIST, “Guidelines for Media Sanitization, Special Publication 800-88”, Internet: https://nvlpubs:nist:gov/nistpubs/SpecialPublications/NIST:SP:800-88r1:pdf, 2014, [Mar. 29, 2021].

[28] N. I. S. Program, “DoD 5220.22-M, Operating Manual”, Internet: https://www:esd:whs:mil/Portals/54/Documents/DD/issuances/dodm/522022M:pdf, 2006, [Mar. 29, 2021].