SENTIMENT AND BEHAVIORAL ANALYSIS IN EDISCOVERY

_____

A Dissertation

Presented to

The Faculty of the Department of Computer Science

Sam Houston State University

_____

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

_____

by

Sundar Krishnan

August, 2022

SENTIMENT AND BEHAVIORAL ANALYSIS IN EDISCOVERY

by

Sundar Krishnan

_____

APPROVED:

Narasimha K. Shashidhar, PhD
Dissertation Chair

Cihan Varol, PhD
Committee Member

ABM Rezbaul Islam, PhD
Committee Member

John Pascarella, PhD
Dean, College of Science and
Engineering Technology

DEDICATION

First and foremost, I would like to praise and thank God, the Almighty, who has granted me countless blessings, knowledge, and this opportunity for me so that I have been finally able to accomplish this dissertation and doctoral program. This dissertation is dedicated to my little daughter Krupa. Your arrival during my dissertation was the most blessed moment of my life. My time spent with you is my favorite. It was never easy on my heart to stay away from you to pursue my dissertation while you slept. To my wife, Nutan. Thank you for being my sounding board, for making time for me to pursue my dissertation while you took care of our little one. Without your unconditional support, I could not have completed my dissertation and my doctoral program. To my dad and mum who are largely unaware of my doctoral degree but will indeed be happy that I earned it as they never dreamed that I could come this far. I love you both and am very proud to be your child! To my in-laws who have happily supported me in this journey. Many thanks to your understanding and well-wishes. To all my teachers in my life: at my kindergarten, school, colleges, and the numerous learned colleagues at work. Many thanks for unconditionally imparting your knowledge to me. Thank you and I am forever grateful to all of you who have taught me.

ABSTRACT

Krishnan, Sundar, *Sentiment and behavioral analysis in ediscovery.* Doctor of Philosophy (Cyber and Digital Forensics), August 2022, Sam Houston State University, Huntsville, Texas.

A suspect or person-of-interest during legal case review or forensic evidence review can exhibit signs of their individual personality through the digital evidence collected for the case. Such personality traits of interest can be analytically harvested for case investigators or case reviewers. However, manual review of evidence for such flags can take time and contribute to increased costs. This study focuses on certain use-case scenarios of behavior and sentiment analysis as a critical requirement for a legal case's success. This study aims to quicken the review and analysis phase and offers a software prototype as a proof-of-concept. The study starts with the build and storage of Electronic Stored Information (ESI) datasets for three separate fictitious legal cases using publicly available data such as emails, Facebook posts, tweets, text messages and a few custom MS Word documents. The next step of this study leverages statistical algorithms and automation to propose approaches towards identifying human sentiments, behavior such as, evidence of financial fraud behavior, and evidence of sexual harassment behavior of a suspect or person-of-interest from the case ESI. The last stage of the study automates these approaches via a custom software and presents a user interface for eDiscovery teams and digital forensic investigators.

KEY WORDS: Ediscovery, Digital forensics, Machine learning, Data mining, Security, Financial fraud, Sentiment analysis, Sexual harassment, Securities fraud, Legal analytics

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentor and dissertation chair, Dr. Narasimha K. Shashidhar, for always challenging me, providing invaluable advice, and for your unconditional support through my doctoral program. Your encouragement and guidance have helped me become a better writer, a critical evaluator, and an improved scholar. Thank you for believing in me, and I look forward to your continued mentorship and our future collaborations. I want to thank my dissertation committee members, Dr. Cihan Varol and Dr. ABM Rezbaul Islam, for all your thoughtful and constructive feedback and support throughout my dissertation. I thank Dr. Varol for challenging me to evaluate abstract concepts and apply them to my research. I thank Dr. Islam for guiding me in analytics and machine learning during my dissertation. I would like to thank the Department of Computer Science faculty for their encouragement in conducting various research experiments and for the valuable knowledge they have imparted. I would like to also thank the staff of the Department of Computer Science for their invaluable assistance in completing various forms and paperwork. I want to thank my entire family for their unconditional love, support, and encouragement during this doctoral program.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER I

INTRODUCTION

Digital evidence can be used to prosecute both civil and criminal cases wherein the evidence sources can involve multiple electronic devices that may contribute to vast volumes of evidence. Sifting through this voluminous pile of evidence in a civil case is known as eDiscovery or electronic discovery. Discovery is a legal process that governs the right to obtain and the obligation on legal parties to produce non-privileged matter relevant to any party's claims or defenses during litigation in state and federal courts. Although the digital age dawned in the late 1990s, "eDiscovery" did not officially become part of legal parlance until 2006 due to the sheer volume of ESI (Electronically Stored Information), including forensically extracted evidence generated by organizations as part of the civil litigation.

There have been predictions that the Global Datasphere will grow from 175 Zettabytes by 2025 (a Zettabyte is a trillion gigabytes) [1], and annual global IP traffic will reach 4.8 Zettabytes per year by 2022 [2]. The number of devices connected to IP networks is predicted to be more than three times the global population by 2022 [2]. Compounding these data volume challenges is the global regulatory environment's growing complexity, such as anti-bribery, anti-corruption enforcement acts, and foreign data privacy laws. In this highly digitized, regulated, and litigious age, a robust electronic discovery program is essential for organizations to avoid litigation inefficiencies, increased costs, increased risks, and possible allegations of incompetence or non-compliance. With the series of opinions authored by Judge Shira Scheindlin in Zubulake v. UBS Warburg [3] heard between 2003 and 2005 in the United States and the thereafter revisions to the Federal

Rules of Civil Procedure (FRCP) [4], a new industry within legal practice appeared known as eDiscovery. Prior to this, discovery in litigation had existed for many years, with opposing parties and their lawyers making requests to the presiding judge/arbitrator to exchange documents relevant to the case. EDiscovery transformed this traditional process from the paper-based pre-Internet world of discovery to a whole series of defined rules and decisions related to how to identify, collect, preserve, analyze, review, produce and present electronically stored information (ESI) or digital case evidence. Such digital evidence exists in a wide range of media and formats such as routine office communications (word processing, spreadsheet files, emails, etc.) to photographs, blog postings, videos, personal emails, social media, and website data. Recent court cases have been peppered with electronically stored information posted on social media sites such as Facebook and text/video messages on mobile devices stemming from applications such as WhatsApp, YouTube, Facebook, Tinder, Twitter TikTok and Snapchat. eDiscovery effort is now no longer limited to focusing on the laws of the land but has now crossed into the realm of technological, logistical, privacy, security, and ethical issues. The Federal Rules of Civil Procedure (FRCP) have been continuously revised with the goal of making the eDiscovery process more efficient, less burdensome, and less costly. This study is particularly timely, given that electronic evidence is increasingly created from our daily life activities and extracting meaningful information to win legal arguments from such voluminous data is increasingly a challenge.

Electronic Discovery

Electronic discovery (also known as e-discovery, eDiscovery, eDiscovery, or e-Discovery) is the electronic aspect of identifying, collecting, analyzing and producing

electronically stored information (ESI) in response to a request for production in a lawsuit or investigation[5]. The process includes collecting, processing, and classifying large corpora of ESI across networks and the Cloud. It has spawned a global support industry with annual spending revenue estimated at 11 billion dollars in 2019 with an estimated growth of approximately 12.93% to $20.63B in 2024 [6]. The same report also finds about 63% of worldwide eDiscovery software and services spending in 2019 occurring in the U.S. Examples of ESI artifacts includes, but is not limited to, emails, documents, chats, presentations, databases, voicemail, audio and video files, social media, and web sites. The processes and technologies around eDiscovery are often complex because of the sheer volume of electronic data produced, analyzed, and stored. The eDiscovery (electronic discovery) process can be outlined by the EDRM model [7], which provides a conceptual view of the eDiscovery process. Table 1 describes the steps in the EDRM process. EDiscovery, is often managed by technology specialists, whose expertise is usually in managed document reviews. An army of such personnel can reduce the stress of high stakes complex litigation. They can help craft the final storytelling of the case, the trial strategies, all future arguments, motions, and depositions, ensuring a robust and comprehensive trial preparedness, proving beyond a reasonable doubt the merits of the case. EDRM stages can be described as below. The EDRM workflow can be an iterative process and can cycle back to the previous stage numerous times when trying to refine results from each stage.

Electronic discovery in any legal case primarily revolves around the review and analysis process. This step of the ERDM is the most laborious and time-consuming. Review involves in-depth analysis of collected documents to determine which ones are relevant to the case, which ones are not relevant, and which ones contain privileged or other

protected information. Reviews are usually performed in iterations and are highly skill and tool dependent. Reviews can also happen in the early stages of the EDRM process, such as the collection stage, and are known as Early Case Assessment (ECA) [8].

TABLE 1
eDiscovery Stages [EDRM]

Stage	Description
Information Governance	Following Industry best practices around Information Governance to mitigate risk and cost should eDiscovery become an issue, from the initial creation of ESI through its final disposition. This pertains to Information governance to both stakeholders and legal entities themselves who practice eDiscovery. All organizations/businesses that deal with electronic data should follow Industry best practices when it comes to Information Governance.
Identification	Locating potential sources of ESI and determining its scope.
Preservation	Ensuring that ESI is protected against inappropriate alteration or destruction.
Collection	Gathering ESI for further use in the e-discovery process (Processing, review, etc.).
Processing	Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review and analysis.
Review	Evaluating ESI for relevance and privilege.
Analysis	Evaluating ESI for content and context.
Production	Delivering ESI to others in appropriate forms and using appropriate delivery mechanisms
Presentation	Displaying ESI before audiences (at depositions, hearings, trials, etc.)

Case Data Volume

The opportunities and challenges presented by today's growing flood of data can be seen via a wide range of products, technologies, and systems, from blockchain technology to digital health, and especially today's high-profile autonomous vehicle initiatives. Individuals and major corporations are processing and storing large volumes of data (from Gigabytes to sometimes Terabytes). Such an aggregation of data from traditional sources (structured data), sensory sources (metadata), and social media (social

data) is also known as “Big data” [9]. During litigation, such volumes of data could be potential evidence that can quickly drive-up litigation costs as human resources are majorly employed in their eDiscovery life cycle. Big data can mean big expenses if proactive measures are not taken around information/data governance. A promising application for big data analysis is fast, high-performing data analytics and data mining that can substantially reduce the time and cost of preparing for a case. In fact, at the root of any eDiscovery effort or process is the ability to smartly identify, efficiently collect, index and analyze big data [10]. Thus, eDiscovery teams are often tasked with processing raw unstructured data to structured data for their consumption and analysis. This processing/synthesis of unstructured to structured data ultimately helps with legal oral arguments and case outcomes.

Technology Assisted Review

A major slice of eDiscovery costs is incurred during the review process. Historically, many eDiscovery solutions/tools have focused on improving collection efficiency and reducing data review effort for long. However, as digital evidence can quickly swell for a case, the costs in forensically extracting data from evidence and then culling this data to arrive at the few select documents critical for legal arguments can be a costly exercise. In 2012, EDRM proposed Technology-Assisted Review (TAR) [11] and has since gained popularity with the legal industry as an essential tool during eDiscovery. The TAR framework (also known as predictive coding) refers to a document review approach that leverages computer algorithms to identify and tag potential documents based on keywords and metadata. Simple Active Learning (SAL) or Simple Passive Learning (SPL) was initially the focus of TAR framework. A second generation of Technology-

Assisted Review (TAR 2.0) however focuses on Continuous Active Learning (CAL). This continuous approach to learning 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 [12]. As a result, the review progresses by re-ranking the entire data set with each new batch of data in a non-iterative and continuously improving implementation of TAR [12]. Thus, 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 [13]. Thus, TAR 2.0 has many advantages over TAR 1.0 and has garnered favor with judges familiar with its benefits [14] while also downplayed by judges refusing to compel parties to apply TAR [15]. 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. In short TAR 2.0 paves the way for leveraging analytics, data mining and automation for case preparation.

Analytics – Legal and Data

The technology umbrella of TAR incorporates analytical techniques such as Machine Learning (ML) (supervised or un-supervised), Artificial Intelligence (AI), Deep Learning, Neural Networks and Statistical approaches. Of recent, with the increasing focus on machine-learning and artificial intelligence across multiple industries, these statistically driven techniques have gained popularity with legal firms and eDiscovery solutions vendors. The application of these techniques in the legal industry has helped coin the term “Legal Analytics” but is otherwise known as “Data Analytics” or “Forensic Evidence

Mining". Few use cases for such analytics involve expedite the organization and prioritization of document collection, improve eDiscovery workflow efficiency, motion forecasting, minimize review efforts, legal strategy, billing optimization, settlement award, resource management, and financial operations. In a digital forensic setting, forensic evidence data can be mined to extract knowledge such as suspect's profile, matching pictures, face recognition and predict intents. These techniques help save costs and reduce time in helping to identify relevant data. In certain scenarios, supervised learning is employed for predictions if historical data of good quality is available else, other techniques such as clustering of data or unsupervised learning are employed for predictions. These analytical techniques greatly focus on quality of raw data to find meaningful patterns, predict the future, and give insights into data. Often raw evidence data is likely to be imperfect, noisy, inconsistent, and sometimes redundant, making it unfit for direct analysis. The process of carefully cleaning/transforming raw evidence data into a consumable state for analytical experiments (data preprocessing) is often a prerequisite.

The courts are the ultimate proving ground in accepting and validating any analytical techniques used for legal arguments. While legal/data analytics techniques help lawyers make data-driven decisions on which to build their legal strategies [16], they tend to have limitations as they clearly do not run the investigation but, merely assist in speeding up the overall process. Also, employing analytics in an investigation or eDiscovery is still in its nascent stages as of today because, if used in many waypoints along the overall process, the courts and opposing counsel can start to doubt is underlying logic as analytics is a complicated mathematical and statistical process by itself.

Suspect Sentiments and Behavior Analysis

A specialization in the processing of unstructured to structured legal data from the evidence pile is in the analysis of the actors/subjects (human) behavior. To understand the behavior of the people (actors/subjects) involved in the case timeline, their sentiments, events in their life, actions that they take, their interactions with others, and their motives, play an important role. Human sentiment can be defined as an attitude, thought, emotion, opinion, or judgment prompted by feeling intended to be conveyed by words, acts, or gestures [17], [18]. Many legal cases rely on understanding human behavior to prove motive, identify opportunities and means that were employed. Their behavior over the case timelines can help outline their character, thus contributing to character evidence that can be admitted under Federal Rules of Evidence Rule 404 [19]. According to a litigation forecast in 2018, white collar crime, government contracts are among a few types of legal cases that will flood the courts [20]. A common type of crime is financial fraud, which is common to many legal cases both civil and criminal. Another type of crime is that of sexual harassment which is widely prevalent in workplaces. Federal, state, and local employment discrimination laws provide a range of legal remedies to victims of sexual harassment. Both these types of crime involve profiling human behavior.

Technology driven tools for automation of forensic investigations of fraud are few and often proprietary to the forensic investigation teams. The bulk of securities fraud investigations involves many painstaking hours of combing evidence. Financial evidence data specific forensic tools may exist, but there is a lack of tools that mine non-finance data for clues of financial fraud. With fraudsters leveraging the Internet for social media platforms, finance related discussion forums, smartphones, etc., evidence of fraud has now

moved away from traditional financial data. This also calls for exacting such evidence from networks, smartphones and computers, thereby involving digital forensic professionals. Preparing clues for prosecution may later involve eDiscovery professionals and paralegals.

Sexual harassment can be categorized into three types: verbal/physical, written and visual depending on the setting/scenario. Written is probably the most common and obvious at workplaces and over the Internet. In daily life, verbal sexual harassment can occur in public settings, on dates or at parties. Visual sexual harassment usually tends to follow verbal or written. Few written sexual harassment examples are emails with offensive jokes, requests for dates, comments on clothing, asking for sexual favors, and graphics with a sexual hint, about race/religion, making derogatory comments about someone's disability or age. While perusing existing literature on the detection of sexual harassment using machine learning and neural network techniques, there was a lack of detection using human intent. Sexual harassment cases, irrespective of type, can have intents such as persuasion, display of power, abuse, victim humiliation, and unwelcome gestures. These intents with sexual overtones can further cement a case of sexual harassment. However, extracting such flags in written conversation can take time due to the volumes of electronic data churned out these days by people. In this study, multiple analytical techniques are employed to propose an approach in identifying sexual harassment indicators leveraging the perpetrator's intent from textual evidential data. This approach can be leveraged by investigating teams who have no adequate labeled data for model learning.

Electronic Discovery Software

Ediscovery software and techniques empowers legal teams to manage litigation scope, response, investigations, and information requests. Often the goals of eDiscovery are to lower e-discovery costs, reduce risk and improve litigation resolution speed. Of recent, eDiscovery product vendors have started to focus on Machine Learning and Artificial Intelligence to improve discovery speeds and accuracy. Development of a custom software/tool catering to evidence mining, AI, and Natural Language Processing (NLP) needs careful planning on design, data privacy and scalability. Few software have made it to the market such as Relativity's Technology Assisted Review [21] and Fronteo's Artificial Intelligence engine "KIBIT" [22]. In this study, an end goal is to design, develop and demonstrate a custom software that automates the handling of case evidence and leverages analytics to predict sentiments of case suspects, indicators of financial fraud and sexual harassment of suspects while pointing to their evidence sources within the case ESI.

Problem Statement

The aim of this proposed dissertation research is to assemble datasets, identify approaches/solutions and develop an application that addresses the below problems.

1. Limited availability of public datasets for eDiscovery ESI analysis that mimics a real-world legal case. There have been some attempts at using the highly redacted Enron dataset [23], but, this dataset does not serve our purpose as we need actor/subject metadata, email metadata etc. that has been redacted. There is some effort in creating fictional public datasets, but they remain limited to the research teams.

2. Limited research has been undertaken to document the conversion process of unstructured raw data of the legal caseload into a structured format. There have been some research attempts at documenting the conversion process but not in a legal context.

3. Ascertaining sentiments of people is a highly researched field but building a sentiment profile of actors/subjects against timelines of the legal case has not yet been undertaken. Such profiling can significantly contribute to the suspect's character profile and legal arguments of a case.

4. Identifying fraudulent financial behavior of actors/subjects/suspects against timelines of the legal case can greatly assist in quickly isolating key case documents for detailed reviews. Significant research has been undertaken in identifying and predicting financial fraud, but they have been limited to credit card data, financial data, or financial statements. This behavioral analysis approach from a legal case ESI can be considered as a novel approach and can significantly reduce the time (person-hours) required to isolate case documents when such behavior is key for legal arguments.

5. Identifying sexual harassment behavior of actors/subjects/suspects against the timelines of the legal case can greatly assist in quickly isolating key case documents for detailed reviews. Limited research has been undertaken in this field. This behavioral analysis from a legal case ESI can be considered as a novel approach and can significantly reduce the time (person-hours) required to isolate case documents when such behavior is key for legal arguments.

6. Limited Graphical User Interface (GUI) software application/tools exist in the industry that allow sentiment and behavior analysis of suspects from legal case evidence load. Such

a software can help serve as a single platform to upload a case, provide case background indicators, and output sentiment and behavior results.

Motivation

Behavioral profiling of an individual often involves identifying and studying their movements, coupled with analyzing patterns in their sentiments and behavior. Profiling an individual using analytical techniques from a caseload (ESI) can greatly assist with preparing case analysis and legal arguments during eDiscovery. Such profiling can also be used in criminal investigations by law enforcement agencies to identify likely suspects from digital evidence. Litigation involving insurance, loans, antitrust, banks, banking, forfeiture, securities, commodities, exchanges, etc. are candidates for financial fraud scenario investigations wherein eDiscovery could be enormous in both time and effort. Legal cases in the U.S. District Courts involving potential financial fraud investigations since 2014 show a steady market for eDiscovery efforts [24], [25]. According to a Global Banking Fraud Survey [26], retail banks experienced increases in total fraud value and volume in 2019. Increased fraud scenarios included identity theft, account takeover (ATO), card not present, and authorized push payment scams. Financial fraud investigations can lead to civil or criminal cases and can involve institutions and governments across borders. Sexual harassment at workplace - be it quid pro quo or a hostile work environment, is a constant battle for victims and employers. Also, sexual harassment behavior can manifest into cyber-bullying, predatory tactics or child trafficking making it a dire area to focus on. Currently, eDiscovery software tools do not integrate behavior analysis functionality within their existing products. Ediscovery software vendors have indicated this area as a future functionality offering. Certain aspects of behavior analysis also touch upon

physiological analysis, risk prediction, and criminal profiling. With additional background information of subjects/actors of the legal case, information can be further enriched. This enriched information implemented by a custom algorithm that factors certain patterns and flags gleaned from the case data can lead to localizing and predicting behavior of subjects/ actors. However, there has not been enough authoritative research especially in this field during eDiscovery of a legal case. Until now, most of financial fraud detection research is limited to credit card data, financial spreadsheets, and financial statements. Little to no research exists in detecting or predicting fraudulent financial behavior among a legal case actors/subject. Similarly, little research has been undertaken in detecting or predicting sexual harassment behavior among a legal case actors/subject. Detecting or predicting such human behavior can establish an affirmative link between actors in their motive by leveraging an opportunity and utilizing means to commit fraud or harassment. In this study, a series of research goals were identified starting with assembling a legal case ESI (evidence repository), identifying human actor's sentiment and behavioral aspects, profiling a human actor from the evidence pile for his/her sentiments and possible fraudulent financial and sexual harassment behavior based on the actor's activity and timelines. Lastly, a custom Graphical User Interface (GUI) software was developed that automated this process. This application/tool inputs a legal case ESI and outputs the sentiment and behavior of key human actors in a timeline based visual format. This study serves to be a significant resource for eDiscovery professionals, legal support teams, law enforcement, digital forensic investigators, and the research community.

Significance of Study

This study contributed to the development of a custom software application/tool that can ingest a caseload ESI volume of files from a stored drive, list human actors in the ESI and display their sentiment and behavior over a timeline. Such profiling of suspects/actors through an automated tool greatly helps legal, eDiscovery and forensic teams in their evidence review process and case argument preparation. Table 2 shows benefits of this dissertation research. The application/tool caters to offender/actor profiling and assisted in the following ways.

1. Provide useful investigative information on offenders/actors.
2. Give investigators some information to work with.
3. Identify personal characteristics to help solve investigations.
4. Assist with plotting offender's signature.
5. Assists with understanding the motives of a crime.
6. Provide a timeline/linkage analysis of suspects/actors in the case.

TABLE 2

Benefits of the custom application/tool towards challenges in EDRM [27] stages

EDRM Stage	Solution to eDiscovery industry issues/challenges by the application/tool
Information Governance	1. Information Governance Model - The application/tool reporting against a particular ESI storage can help improve overall governance framework of a legal entity across other customer ESI storage by better securing information assets, managing risks, tuning disaster recovery procedures, preparing for contingency and business

(continued)

	continuity. The application/tool reporting can also indirectly help fine-tune the governance model by fine-tuning policies, procedures. 2. Data Privacy - The application/tool reporting identifies electronic assets with the ESI repository that would need adequate privacy controls.
Identification	The application/tool reporting identifies broken links or gaps that will allow for a re-identification.
Preservation	The application/tool reporting identifies broken links or gaps that will need revisit of preservation controls and a process review. It can also help confirm/extend existing legal holds.
Collection	If collection process was inadequate, the application/tool reporting outcomes identifies broken links or gaps that will need revisit of collection controls and a process review. It also helps in identify collection errors and thus allow tuning of the process.
Processing	The application/tool reporting outcomes identifies broken links or gaps that will need a revisit of processing controls, redaction techniques and process review. It can also help identify processing errors and thus allow tuning of the overall process
Review	The application/tool reporting outcomes identifies broken links or gaps that will need collection controls and a process review. It can also help identify review errors and thus allow tuning of the process.

(continued)

Analysis	1. The direct and considerable impact of the application/tool will be in the analysis phase. The solution greatly reduces time taken by the legal team to analyze the ESI repository there by saving cost. 2. The application/tool greatly reduces manual errors by achieving a high degree of accuracy. Thus, rework effort can be minimized leading to reduced cost. 3. The application/tool allows for unsupervised learning re-runs if needed. 4. The application/tool can scale with ESI repository volume. 5. The custom software/tool can identify sensitive data across the ESI repository thereby allowing for a review of redaction, privacy, security controls on identified data.
Production	1. The application/tool can assist in production process by reducing time and selectively identifying the needed data for production. 2. The application/tool can identify sensitive data across the ESI repository thereby allowing for a review of redaction, privacy, security controls on identified data during production.
Presentation	The application/tool greatly assists the legal team in presentation due to extensive reporting. 2. The application/tool identifies sensitive data across the ESI repository thereby allowing for a review of redaction, privacy, security controls on identified data during presentation.

(continued)

	3. The application/tool allows for a timeline visualization to assist with the presentation process.

Limitations

The purpose of this dissertation was to profile human sentiments, behavior patterns from evidence of a legal case. This dissertation only addresses legal case (ESI) artifacts (files) in English (US) language with the provision to scale across other languages. Initial ESI repository was limited to .pst, .csv (social media data) and MS Word files. Initial results were limited to semi-contextual analysis and progress into fully contextual as part of future work. All social media data of the given legal case (ESI) were assumed to already forensically obtained for the sake of ease. As a use-case for fraudulent financial behavior, this study only focused on "Insider Fraud in an organization" and general "pump and dump scenario" from textual communications of suspect(s) in the legal case. This selection was due to ease and simplicity. This study focused on general sexual harassment indicators in textual communications of suspect(s) in the legal case. Again, this selection was due to ease and simplicity. The custom GUI platform/tool/software functionality was limited to only automating this study's goals.

Definition of Terms

Analytics: Systematic computation and analysis of data for meaningful patterns by leveraging techniques such as Machine Learning (ML) (supervised or un-supervised), Deep Learning, Neural Networks, and other statistical approaches.

Evidence Data Analytics: When analytical techniques are used against case evidence in scenarios such as to preparing for wining case arguments, prediction of

criminal intent, suspect profiling, matching a suspect face against digital pictures, identify a deep-fake.

Legal Analytics: When analytical techniques are used in scenarios such as predicting legal costs, predicting case timelines, knowledge mining of past court opinions on similar cases.

Financial Fraud: A fraudulent activity that occurs when someone takes money or other assets from through deception or criminal activity.

Sexual Harassment: Unwelcome sexual advances, requests for sexual favors, and other verbal or physical remarks in a workplace or other professional or social situation.

Organization of the Study

This study consists of eight chapters. Chapter I includes the statement of the problem, motivation, significance, research questions, and limitation. Chapter II includes review of literature involving preparing for eDiscovery, identification of case evidence, collection of case evidence Data, existing datasets for eDiscovery (legal) analytics, database platforms and data conversion techniques, keyword searches and natural language processing, identification of human actors from a legal case ESI, sentiment analysis in a legal case ESI, Behavior Analysis around Financial Fraud Behavior and sexual Harassment Behavior in a legal case ESI and existing eDiscovery Tools focusing on Behavior and Sentiment Analysis. Chapter III contains the data collection methods, sources, technology selection, and data preprocessing steps. Chapter IV discusses the sentiment analysis of suspects in a case investigation, methodology, analysis, results, and summary of this chapter. Chapter V discusses the analytical approaches to detection of financial fraud of case suspects such as insider trading and pump and dump schemes, analysis, results, and a

summary of the chapter. Chapter VI discusses the analytical approaches to detection of sexual harassment indicators of case suspects analysis, results, and a summary of the chapter. Chapter VII contains the design and operational functionality details of the custom forensic software/application/tool prototype to help case investigators in automation of sentiments analysis, financial fraud detection and sexual harassment detection (Chapters IV, V and VI), best practices and challenges in custom forensic software development when implementing analytics, followed by a summary of the chapter. A conclusion of this dissertation and future work are presented in Chapter VIII.

CHAPTER II

LITERATURE REVIEW

In the current digital world, data continues to grow exponentially, given the rampant use of computers, readily available Internet on smartphones, IoT devices, Smart devices, Cloud etc. In a legal setting, this means finding hidden needles in a haystack due to the enormous data that legal minds must sift through to find the relevant data to the case. K&L Gates continually updates a searchable database containing more than 3,000 electronic discovery cases collected from state and federal jurisdictions around the United States [28]. In a 2019 survey of 102 Law Firms [29], about 86% of them pass on actual costs related to discovery directly through to the client, with their biggest challenge being difficulty in predicting data size. Pace et al. [30] gathered costs for 57 large-volume e-discovery productions and conclude that 73 cents of every dollar spent on electronic production was spent on the review stage of the EDRM model [31]. This suggests that costs associated with large-scale document reviews dominate total production expenditures. They also conclude that promising alternative technologies available today for large-scale reviews use predictive coding and categorization strategies to rank electronic documents by their relevance/privileged likelihood. These alternatives can be accomplished using statistical technique driven approaches and artificial intelligence methods such as machine learning and neural networks. For these alternatives to work accurately and flawlessly, algorithms must be employed against electronic datasets that mimic a legal case electronic information.

Preparing for eDiscovery

These days, enterprises need to be prepared for eDiscovery from the first day of business as they cannot predict when litigation can occur, and this preparation is part of the Information Governance stage of the EDRM framework [31]. In U.S. courts, legal precedent requires that potentially relevant information must be preserved at the instant a party "reasonably anticipates" litigation [32]. The event or occurrence that causes the party to begin preserving information (digital or otherwise) is referred to as the "trigger" or "triggering event". A triggering event can be subpoenas or cease-and-desist letters, the threat of litigation, regulations, Preservation orders, etc. This process is known as a "hold" or "legal hold" and triggers preservation to avoid spoliation. In the U.S., 49 states have adopted statutes and court rules addressing the discovery of electronically stored information [33]. Chris Delgado [34] discusses the preparedness and compliance for eDiscovery by providing a bibliographic guide of various eDiscovery tools, legal sources, Federal Rules of Civil Procedure rules, statutes, RSS feeds, etc. that enable the E-Discovery process. Business Enterprises, schools, non-profit organizations, etc. should be prepared for litigation and preparation starts with the governance. Entities should have their electronic systems and data to mitigate risk & expenses should e-discovery become an issue, from the initial creation of ESI through its final disposition [35]. This stage in the eDiscovery process is well discussed on blogs and whitepapers of law firms, but not on academic papers especially when applying Artificial Intelligence (AI) for subsequent analysis. This gap in research is proposed to be discussed through thesis goal #1 and on proposed research papers when assembling legal case evidence (ESI) dataset(s).

Identification of Case Evidence

Identifying case evidence is part of the EDRM framework [7]. It involves applying holds and identifying potential evidence - be it across an organization's infrastructure or the personal lives of individuals linked to the legal case. eDiscovery professionals are trained to assist in identification. At an enterprise, standard operating procedures and policies help the information technology teams to assist their legal counsel with a legal hold on electronic documents. Negangra et al. [36] prescribe an instructional case for students using Enron data [37] and the EDRM lifecycle to introduce how digital evidence is incorporated into a forensic accounting investigation and challenges them to learn electronic discovery tools and techniques by thinking before digitizing. Bernier [38] discusses technology-neutral measures for attorneys to identify search protocols, plot initial course, measure progress by measuring the initial criteria's accuracy, determine what to review by tuning search protocols, sampling, making mid-course corrections by modifying the criteria all when performing eDiscovery. Hyman et al. [39] proposed a design for retrieval of artifacts using a bag of words (BOW) approach for terms (based on initial terms, synonyms, and slang) and a standard deviation method for assigning weights. In another article, Hyman et al. [40] discuss the information retrieval (IR) problem of balancing recall with precision in electronic document extraction. They conduct behavioral experiments to examine the IR constructs of uncertainty, context and relevance while proposing a new process model for context learning by leveraging explicit knowledge to discover implicit knowledge within a corpus of documents. Identification of case evidence during eDiscovery is to be discussed through thesis goal #1 when assembling case evidence (ESI) dataset(s).

Collection of Case Evidence Data

In the EDRM framework [7], the collection of evidence follows their identification. With technological advances each day, coupled with the growing appeal of smart mobile devices, the collection of evidence for a legal case can go beyond traditional computer disks. In outlining common problems of eDiscovery, Hernandez [41] notes that sifting through massive amounts of data is no longer a problem solely for mass tort and class action cases but is a problem for small size cases like a breach of contract. Increasing volumes of data to acquire and analyze directly increases eDiscovery costs and time. The collection of potential digital evidence from various sources can involve implementing forensic techniques. Thus, the collection effort is usually undertaken by skilled forensic professionals and may require multiple sub-disciplines of digital forensics like accounting forensics, IoT forensics, Smartphone forensics, Cloud forensics, etc. The collection of case evidence during eDiscovery is to be discussed through thesis goal #1 when assembling case evidence (ESI) dataset(s).

Datasets for eDiscovery

Few datasets (corpora) exist for Machine Learning in eDiscovery other than the decades-old, sanitized Enron Corporation legal case dataset [42] that stems from one of the largest and most complex civil fraud trials in U.S. history. This has been the go-to set for all eDiscovery product vendors and is freely available in the public domain post de-duplication at the custodian level. In recent times, claims that it's no longer a representative test data set for eDiscovery solution testing (processing) has gained traction [43]. However, this dataset is still popular in academic research. Noever [44] applies machine learning and uses the Enron Corporation dataset to identify persons of interest (POI) with an accuracy

of 95.7% and discover 50,000 previously unreported instances of PII and flag legally responsive emails with a 99% accuracy. The author also compares accuracy against execution times for dozens of algorithms and tracks three years of primary topics and sentiment across over 10,000 unique people before, during and after the onset of the Enron corporate crisis. Cori et al. [45] articulate that recent debates and court decisions have focused more on electronically stored information posted on social media sites such as Facebook as well as more informal and transient communications involving text messages such as WhatsApp and Snapchat. Given the rise in popularity of Apps such as Facebook, Instagram and WhatsApp, and the shift to Internet-connected smartphones, tablets, and wearable mobile devices in general, it's just a matter of time for data from these Apps to soon flood eDiscovery cases. However, independent datasets of emails, Facebook posts, MS-Word documents, Social media chats, etc., do exist that could potentially be combined to create a representative dataset of a legal case. This research proposes to combine such discrete datasets for various experiments to augment for the sole availability of the Enron Corporation dataset. This research aims to create three different synthetic datasets comprising publicly available data from emails, Facebook, Twitter, WhatsApp, phone SMS/Texts, and custom MS Word Documents. Each dataset is proposed to mimic a legal case scenario and have a random arrangement of data among a defined set of subjects/actors (suspects).

Sentiment Analysis of Case Suspects

Traditionally, emotional analysis, sentiment analysis and behavioral analysis belong to the psychology, criminal, military, and medical domains. With recent advancements in Machine Learning and Neural Networks, these topics have now been

researched from a digital point of view. These areas now have been leveraged in areas such as the study of employees for risky behavior, chat room actors profiling, understanding the users of the dark web, crowd behavior analysis, stock market analysis and botnet detection. Predicting emotions using Machine Learning algorithms has been well explored. Calix et al. [46] train a model to predict an actor's levels of emotion magnitude prediction in text and speech by comparing linear and non-linear regression techniques. Their results have shown that non-linear regression models based on Support Vector Regression (SVR) using a Radial Basis Function (RBF) kernel provided the most accurate prediction model. Behavior analysis as natural science has been applied in various fields when understanding the behavior of individuals. Tripathi et al. [47] propose a comprehensive survey of current convolution neural network (CNN)-based methods for crowd behavior analysis with emphasis on optimization methods. Leveraging Machine Learning and Artificial intelligence in this field, Romera et al. [48] propose a public dataset to study driver behavior analysis due to the growing safety concerns in vehicles. Another area of behavior analysis is when studying botnets. Garg et al. [49] evaluate various machine learning (ML) algorithms to compare their ability to classify botnet traffic. Haddadi et al. [50] propose a botnet analysis system by implementing two different machine learning algorithms, and Naive Bayes. Shalini et al. [51] perform a comparative analysis of clustering techniques when studying customer behavior. Applied behavior analysis (ABA) also known as behavioral engineering, is a branch of general behavior analysis concerned with applying learning-based empirical techniques to change the behavior of social significance. Foxx et al. [52] state that Applied Behavior Analysis (ABA) uses methods derived from scientifically established principles of behavior that can be effective interventions in

educational and treatment programs for children who have autism. In the criminal world, behavior analysis can be found in criminal or suspect profiling. In eDiscovery, behavior analysis can help better understand the actors in the legal dispute and assist the legal process. Sentiment analysis is contextual mining of data involving natural language processing, text analysis, computational linguistics, and biometrics, which then identifies and extracts subjective information. From the Internet point of view, predicting sentiments of website users can be valuable for user driven action such as displaying advertisements, marketing products and suggesting topics. Liu et al. [53] extend sentiment analysis into opinion mining by analyzing people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes. Xiaomei et al. [54] focus on Twitter and Sina Weibo to investigate how to exploit weak dependency connections as an aspect of social contexts for microblogging sentiment analysis. Hasan et al. [55] adopt a sentiment analyzer using machine learning to analyze twitter accounts. Sentiment analysis has received attention due to the advancements and popularity of machine learning and artificial Intelligence. A literature gap was found to exist regarding analyzing sentiments of case suspects during the eDiscovery and forensic analysis process. Obtaining a sentiment heatmap of suspects in the legal case can help legal minds, forensic and discovery teams arrive at actor sentiments towards co-workers, products, brands or services, businesses etc. The goal #3 of this thesis addresses this gap by designing and deploying a custom software using machine learning and statistical techniques to project an umbrella approach to map sentiments of suspects in the legal case.

Behavior Analysis

According to the New Mexico Association for Behavior Analysis (NMABA), behavior analysis is the scientific study of the principles of learning and behavior [84]. The Association for Behavior Analysis International defines behavior analysis as a natural science that seeks to understand the behavior of individuals [56]. Many research articles cover behavior analysis of malware, consumers, diving of aquatic life, crowd, software architectures, tourists, etc. A large volume of articles is dedicated to Applied Behavior Analysis (ABA), especially in the study of Autism and human diseases. When analyzing and correctly deciphering human behavior, studying their actions and habits for patterns, anomalies, flags, etc. can help create a bigger picture for investigators. Using automation, Machine Learning and by asking investigative questions such as; Why was the action performed?, When was it performed?, What led to this action?, etc. coupled with the categorical trinity of an investigation (motive, means, and opportunity) can help lay a timeline map of human behavior of a legal case. This type of forensics is also known as profiling and overlaps the study of human psychology. Douglas et al. [86] describes this criminal profile generating process as a series of five overlapping stages lead to the last and sixth stage (the goal of apprehension of the offender) as; 1) profiling inputs, 2) decision-process models, 3) crime assessment, 4) the criminal profile, 5) investigation, and 6) apprehension. The Behavioral Analysis Unit (BAU) of the FBI was created to apply behavioral analysis in criminal investigations [57]. This Unit consists of 1) Counterterrorism/threat assessment 2) Crimes against adults, and 3) Crimes against children. The BAU is part of the National Center for the Analysis of Violent Crime (NCAVC). According to the National Board of Forensic Evaluators (NBFE), Forensic

behavioral analysis is a discipline that applies the behavioral sciences to aid law criminal justice agencies, employers, and other organizations in prevention, response, and mitigation of crises, disasters, and critical incidents [58]. In a legal case, mapping the human behavior of all subjects/actors within the legal case can significantly assist in narrowing down on relevant evidential artifacts for further analysis and preparation of case arguments. The below sections describe existing literature in specialized fields of behavioral study in civil ligation namely financial fraud and sexual harassment.

Financial Fraud Behavior of Case Suspect

To solve applied problems and make policy decisions, forecasting or predicting people's behavior is often necessary. Fraud detection has been a topic that has been studied for decades. Computer Assisted Auditing Techniques (CAATs) have been used since the 1960s in audits and have made in-depth data interrogation possible due to their capability of digitally analyze large volumes of data. Two decades ago, Coderre et al. [59] concluded that CAATs could help to ensure that corrupt activity within the organization does not remain undetected. Early on, Persons developed a decision aid using parsimonious models to identify factors associated with fraudulent financial reporting by firms. In his results, the stepwise logistic model produced superior predictive results compared to a naive strategy, which classified all entities as non-fraud firms. Behavior forensics is the application of psychology to financial fraud prevention and detection. Typically, behavioral forensic studies consider employees' characteristics such as greed, lifestyle, personal incentives, etc. that are likely to lead to fraudulent behavior. Rezaee [60] evaluates a fraud audit approach (vulnerability review) in which a forensic auditor can consider whether the company's own structure and processes are likely to lead to the detection and/or reporting

of fraudulent activities. The author also proposes “Gamesmanship Review” in which the forensic auditor assesses the top management teams on their philosophies, attitudes, operating styles, decisions, actions, beliefs, and ethical values. Dunn [61] hypothesized that fraud was a function of structural power, ownership power and control variables. This study showed that excessive power is positively related to illegal corporate behavior. Davidson et al. [62] examine how executives’ behavior outside the workplace as measured by their ownership of luxury goods and prior legal infractions is related to financial reporting risk. The authors predict and find that chief executive officers (CEOs) and chief financial officers (CFOs) with a legal record are more likely to perpetrate fraud but find no relation between executives’ frugality and the propensity to perpetrate fraud. Weatherford [63] highlights few challenges in fraud detection as to coming up with algorithms that can learn and adapt to recognize a great variety of fraud scenarios, identify and predict new scenarios and create systems that work quickly enough to detect fraud activities as they occur. In a survey on financial fraud detection methodologies, Richhariya et al. [64] conclude that only a few approaches for credit card detection are available in public because of security issues. Among approaches, use of neural networks is very popular along with applying machine learning techniques. Predictive supervised algorithms study all known labeled transactions to mathematically uncover how a typical deceptive transaction looks like by assigning risk scores. Decades ago, Ghosh et al. [65] applied a three-layer, feed-forward Radial Basis Function (RBF) neural network for new credit card transactions to churn out a fraud score, in every two hours requiring only two priming passes. The neural network was trained on examples of fraud due to lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud, and NRI (non-received issue) fraud. Syeda et al. [66] apply fuzzy neural

networks on parallel machines to accelerate rule production for customer-specific credit card swindle detection. Barse et al. [67] generate synthetic test data to train the fraud detection system itself in a IP based video-on-demand service. The multi-layer neural network system was then exposed to a set of authentic data to measure parameters such as detection capability and false alarms and compared against synthetic data. Chiu et al. [68] propose a Web services-based collaborative scheme for credit card fraud detection and introduce a Fraud Patterns Mining (FPM) algorithm, customized from apriori, to extract a common format for fraud-only credit card data. Deshmukh et al. [69] implemented an expert system to management fraud transaction data. They illustrated how fuzzy sets can be used to measure red flags on a categorical or interval scale, how different red flags can be combined using fuzzy rules, and how a single measure of the risk of management fraud can be derived. Pervasive supervised algorithms such as Neural networks, Bayesian networks, and decision trees have been applied in fraud detection. Chan et al. [70] use naive Bayes, C4.5, CART, and RIPPER as foundation classifiers and amass to combine them. Their findings resulted in cost savings, reduced loss and enhanced efficiency on credit card transaction fraud through distributed data mining of fraud models. Choi et al. [71] proposed an approach of detecting financial fraud in IoT based on machine learning and compared it with artificial neural networks approach to detect fraud using large amounts of financial data in Korea. Their experimental results showed that machine learning based methods has higher detection efficiency than neural networks. Few articles have grouped together red flags (indicators or patterns or fraudster characteristics) when dealing with white collar crime. Decades ago, Vanasco [72] considered red flags that are likely to be useful in any approach to fraud auditing such as;

1. Looking for analytical symptoms, transactions that seem ’odd’ or out of place.
2. Observing lifestyle or behavioral symptoms such as management’s greed or rich lifestyle
3. Sampling unrepresentative set of items while paying particular attention to transactions that were made outside the usual controlled procedures.

Ramamoorti et al. [73] state that data analytics offers powerful tools and techniques to help deter or quickly detect potential wrongdoing by reaching into volumes of data and identifying anomalies that merit further investigation. The author also states that seven flags can help spot behavioral fraud.

1. Weak code of ethics (willing to engage in dishonest behavior in life).
2. Propensity to work “outside” the system (not following established job procedures and workplace policies).
3. Poor work performance (coupled with the rationalization or justification of a substandard performance, this can be an indicator of disrespect for the organization).
4. Excessive drive to achieve (who desperately try to improve performance or meet certain targets may find it tough to resist the temptation to circumvent fraud controls)
5. Over-protectiveness of data and key documents (Dishonest employees are often reluctant to share information with coworkers or managers).
6. Persistent demoralization (constant dissatisfaction)
7. Being the first one in or the last one out (a non-vacationer)

While adequate research has been taken into detecting financial fraud and few books written by eminent authors, there exists a gap when identifying fraudulent financial behavior indicators using a case suspect's intent to commit financial fraud using evidential case data acquired from emails, social media data and MS Word documents. This gap in research was addressed through this thesis goal #4 in which Natural Language Processing and Machine Learning techniques were leveraged to work with case ESI to predict financial fraud behavior.

Sexual Harassment Behavior of Case Suspect

Sexual harassment is a type of harassment involving the use of explicit or implicit sexual overtones, including the unwelcome or inappropriate promise of rewards in exchange for sexual favors [74]. Sexual harassment includes a range of actions from verbal transgressions to sexual abuse or assault [75]. According to the U.S. Equal Employment Opportunity Commission, harassment can occur in many different social settings such as the workplace, home, school, churches, military, etc. The harassers or victims may be of any gender [76]. In the U.S., sexually harassing someone on the job is against the law and is a form of sex discrimination that violates Title VII of the Civil Rights Act of 1964 [77].

A fertile ground for sexual harassment is at the universities. Ignacio et al. [78] show the scarce presence of technical measures in universities, and offer a set of measures to improve the management of sexual harassment and harassment on the grounds of sex. A lot of victims, particularly women, go through this experience but often do not report them. Bauer et al. [79] built a chatbot based on machine learning and Named Entity Recognition (NER) to assist survivors of sexual harassment to offer them help and increase the incident documentation. The authors were able to achieve a success rate of more than 98% for the

identification of a harassment-or-not case, and around 80% for the specific type of harassment identification. Online social media is another a fertile ground for nefarious activity, specifically sexual harassment, as users take advantage of a virtual environment and use pseudo profiles. The Twitter platform is one such environment where tweets can sometimes linger on the borderline of sexual harassment or jokes. Garrett et al. [80] collected and analyzed tweets from the #WhyIDidntReport Twitter conversation to categorize the reasons why sexual harassment goes unreported by the victims. Using machine learning techniques, they found that hopelessness and helplessness were the most common reasons cited by the victims for not reporting sexual violence incidents. Saeidi et al. [81] employ various machine learning algorithms on Twitter data to predict harassment types with high accuracy. They also showed that, when using TF-IDF vectors, linear and gaussian SVM are the best methods to predict harassment, while Decision Trees and Random Forest better categorize physical and sexual harassment. With the growing accessibility of the Internet and smartphones, sexual harassment and cyberbullying have grown uncontrollably, causing physiological and mental risks to victims. Alawneh et al. [82] propose a machine learning based approach to develop and classify sexual harassment and cyberbullying detection. Their experiments showed that combining Term Frequency Inverse Document Frequency (TF-IDF) with machine learning achieved an 81 % accuracy rate. Basu et al. [83] compare Machine Learning and Deep Learning models to find the most effective model based on contextual clues to predict and classify sexual harassment on social media. While much of the existing literature is focused on classifying social media comments using various machine learning algorithms, this research proposes an

approach to tackle the identification of sexual harassment using perpetrator's intent alongside other risk factors.

In civil litigation, most cases of sexual harassment arise from corporate and personal lives. If an organization or business approaches the problem of sexual harassment with a "one size fits all" solution, chances are high that it may not be protecting some of the most vulnerable members of its workforce. Identifying victims and perpetrators can be complex and in most cases their behavior is to be studied prior and after the alleged incident. There are many other academic studies conducted on sexual harassment related to surveys, studies on other species (non-humans), and psychology. However, not much research has been undertaken to identify harassment behavior from a corpus of digital evidence using suspects intent. This gap in research was addressed through goal #5 of this thesis in which sexual harassment indicators are predicted by leveraging Natural Language Processing and Machine Learning techniques against a case ESI.

Analytical Software Tools for Case Suspect Sentiments and Behavior

Visualization greatly helps any investigation especially when coupled with timeline analysis and profiling of subjects within the scope of the investigation. The existing eDiscovery tools on the market are still largely based on simple string search, pattern searches, RegEx based searches, grouping of files, native file visualizations - all devoid of leveraging Artificial Intelligence (AI), Machine Learning, Statistical analysis and Natural Language Processing. Sathiyanarayanan [84] discusses the challenges and his ongoing research towards an interactive visualization for easy navigation of emails in eDiscovery. The author concludes that developing visual methods, strategies and framework is critical to ease the burden of dealing with voluminous and noisy digital evidence in a legal case.

Digital forensic software and forensic techniques can be used in civil or criminal litigation to extract and analyze evidentiary data from various electronic sources. Artificial intelligence (AI) enabled digital forensic software can boost the analysis efficiency of a digital forensic investigation or in eDiscovery by quickly identifying trends, patterns, anomalies, commonalities, deepfakes, and other traits within the evidence pile. Jarrett et al. [85] conclude that AI-assisted investigations reveal a significant reduction in human mistakes, reducing inquiry time, costs, and wrong outcomes. Mitchell et al. [86] outline a few challenges that face digital forensics when applying AI and finds knowledge representation and ontology as the main challenges. The author finds that the lack of standards hinders the exchange of information between tasks in digital forensic software. Rughani [87] proposed an AI-based digital forensics framework that requires minimum user interaction and does the majority of routine operations by intelligence acquired from training. Digital forensic software leveraging AI can sometimes fail or provide incorrect results. Baggili et al. [88] propose establishing a new discipline of AI Forensics under AI Safety to investigate cases of failure in AI systems. As the digital forensics software industry continues to embrace AI techniques in evidence analysis and presentation, groups of developers and security professionals have started to explore the application of AI reasoning in the Digital Forensics. One such group is the “DigForAsp” (Evidence Analysis via Intelligent Systems and Practices) who acknowledge that no established methodology exists today for digital evidence analysis during an investigation, and experts usually proceed by means of their experience and intuition. Bhatt et al. [89] train an Artificial Neural Network (ANN) and analyze computer RAM along with disk images. They find forensic evidence of certain keywords that are part of the training data.

Legal firms charge their clients-based time taken to semi-manually review volumes of information produced by their tool searches. There is a lot of talk of leveraging AI and few eDiscovery vendors are now offering AI driven add-ons to the existing eDiscovery tools [21], [22]. While the use of analytics in analyzing forensic evidence is steadily increasing, there is little in the way of literature that outlines the development and operations of analytics driven custom forensic software along with its accompanying challenges and opportunities. This gap in research was addressed in the thesis goal #6 in which a custom software was developed to support automation and analytics of goals #3, #4 and #5 of this dissertation. In Chapter VII, the authors propose a custom and functional digital forensic software "Digital Forensic Case Evidence Analytics" (DFCAE) that incorporates analytics and can be used by forensic investigators or eDiscovery professionals in analyzing case evidence for certain clues. The DFCAE software also caters to prior research undertaken by the authors in leveraging AI to mine textual case evidence and is available on GitHub for public academic use [90]. The authors also touch upon the challenges and opportunities faced during the development of this software prior to discussing the highlights of this custom tool.

CHAPTER III

DATASET – COLLECTION AND PREPROCESSING

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. The methodology of this chapter 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.

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. Fig. 1 highlights the sources for deriving Aim/Problem/Objective in legal and forensic analytics.

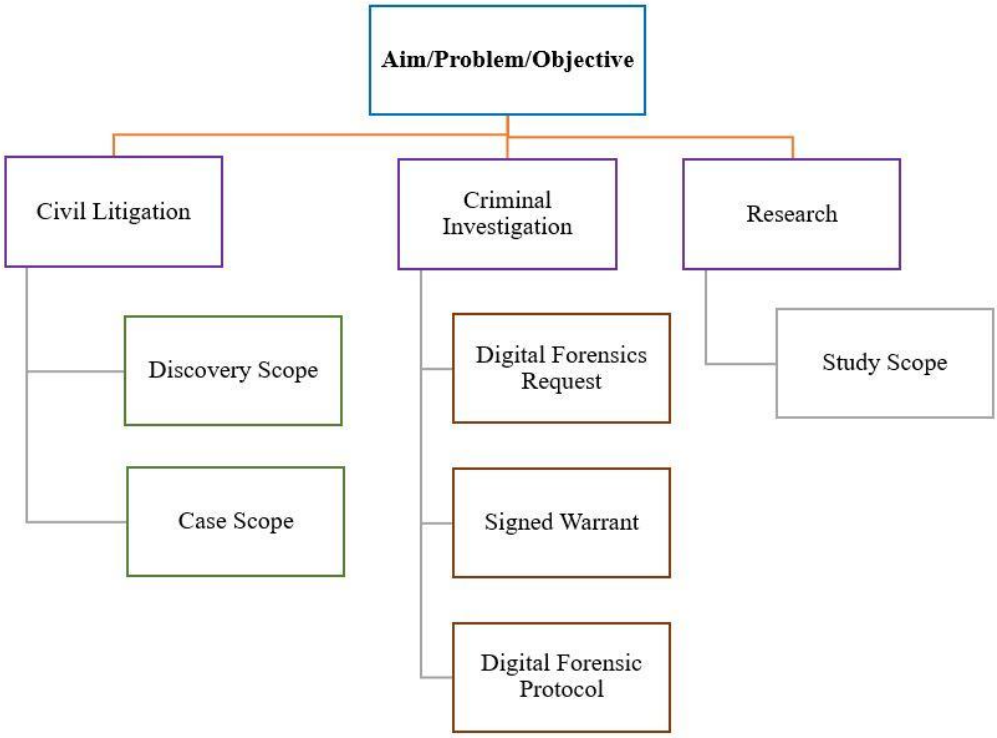

Fig. 1. Sources of analytical aim/problem/objective in legal and forensic analytics.

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 3 outlines few challenges when working with text-based case data. Fig. 2 shows potential issues with raw data of a legal case ESI.

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.

TABLE 3
Common language and text limitations in case evidence data.

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. [91])	"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

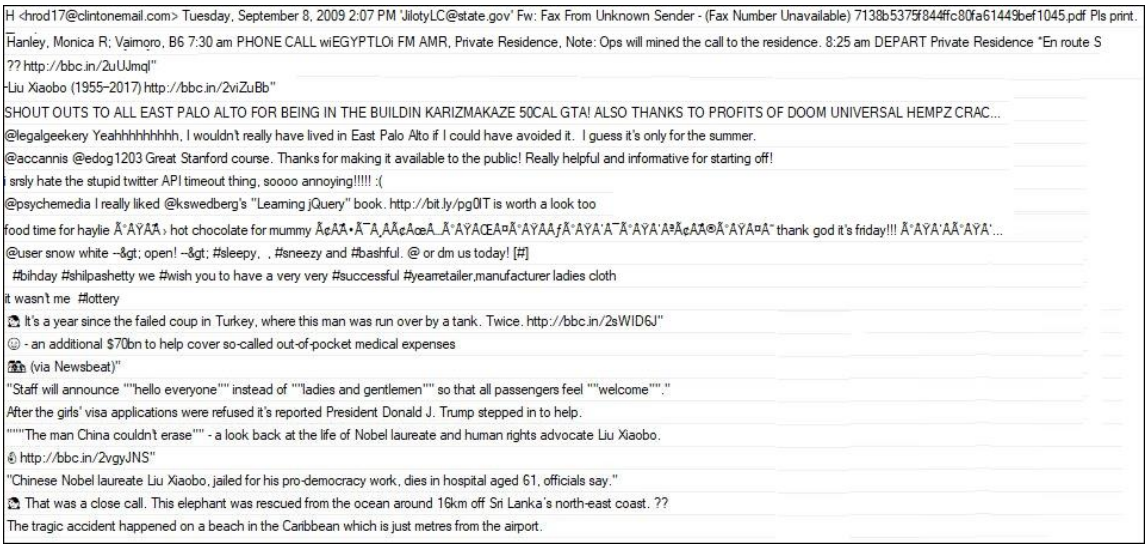
H <hrod17@clintonemail.com> Tuesday, September 8, 2009 2:07 PM 'JilotyLC@state.gov' Fw: Fax From Unknown Sender - (Fax Number Unavailable) 7138b5375f844ffc80fa61449bef1045.pdf Pls print.
Hanley, Monica R; Vaimoro, B6 7:30 am PHONE CALL wiEGYPTLOi FM AMR, Private Residence, Note: Ops will mined the call 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!
srsly hate the stupid twitter API timeout thing, soooo annoying!!!!! :(
@psychemedia I really liked @kswedberg's "Learning jQuery" book. http://bit.ly/pg0lT is worth a look too
food time for haylie Ã°ÃŸÃ‰ hot chocolate for mummy Ã¢Ã‰•Ã¯Ã¸Ã¢Ã¦Ã…Ã°ÃŸÃŒÃƒÃ°ÃŸÃ'Ã¯Ã°ÃŸÃ'Ã‰Ã°ÃŸÃ'Ã¯Ã¢Ã‰Ã®Ã°ÃŸÃ¤Ã¯ thank god it's friday!!! Ã°ÃŸÃ'ÃŠÃ°ÃŸÃ'...
@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.

Fig. 2. Sample raw text in case evidence prior to pre-processing [Note: Evidence can contain garbled characters, Unicode, email addresses, shorthand, slang, URLs, emoji, and hashtags.]

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.

Digital Forensics

There exists an interplay between eDiscovery and digital forensics [92] 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.

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.

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” [93]. 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.

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.

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 [94]. A good model fit is a statistical hypothesis test that determines whether a model accurately approximates the output when given unknown 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; the question is how to spot it and eliminate 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 must be maintained. Bias in analytical experiments can eventually derail a legal case.

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.

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.

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 evidence data through encryption, tokenization, redaction, masking, or de-identification maybe needed for analytical experiments. For example, masking of last names of people in the evidence may be required, or certain geographical location data may need to be obfuscated to protect privacy and identity. If so, such data (features/attributes) may need to be dropped or encoded accordingly during analytical preprocessing. If authorized to use raw data for analytical experiments, care must be taken for storage, distribution and destruction of experiment results, lest, they accidentally expose sensitive data.

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.

Data Integrity of Digital Evidence

Analytical experiments are built on large amounts of data and are increasingly driven by complex feature pipelines with automated workflows that involve data transformations. Data preprocessing steps too can be lengthy when arriving at the best set of data features for the start of analytical experiments. Data integrity of evidence must be protected since indiscriminate processing can terminate or modify 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 evidence data to fit the needs of the analytical experiment should be undertaken carefully and should be well documented.

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.

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.

Backup and Retention

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

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 [95], [96] 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.

Summary

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 analytical approaches such as Machine Learning, Artificial Intelligence, and Deep Learning. In a legal case, digital evidence may be present as digital device data, transmission data, application data, logs, or Internet data. Extracting meaningful data off such evidence data can be voluminous and can burden the analysis and review process during eDiscovery or forensic analysis. 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. When working with evidence data, making verifiable copies, access, logging, along with data storage, backups and destruction is to be planned and approved. This chapter outlines best practices and approach for preprocessing legal data prior to forensic and legal analytical experiments. 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 digital 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.

CHAPTER IV

SENTIMENT ANALYSIS OF CASE SUSPECTS

As real-life forensic investigation evidence or a legal caseload is usually not readily available in public or for academic research, the authors felt the need to build custom caseloads of electronic evidence/Electronic Stored Information (ESI) and later make them available for academic research.

Experiment Design and Methodology

The research experiment revolved around three fictitious legal case ESI or digital forensic investigation evidence (datasets). The experiment was carried out using an Intel(R) Core (TM) i5-3470 CPU @ 3.20GHz 16 GB RAM PC and a 64-bit Windows 10 operating system. Each caseload ESI contained emails (public sources and custom) [97], [23], SMS and publicly available WhatsApp data [98], publicly available Twitter data [99], [100], [101], publicly available Facebook data [102], [103] and a few custom random MS Word files. Each case's ESI was also updated to include random suspect names, as well as a few random posts and tweets highlighting a case/investigation scenario. The labeled dataset for supervised learning was obtained from movie reviews specially selected for sentiment analysis containing 25,000 samples of reviews with binary sentiments [104]. The Fig. 3 highlights the overall design of the experiment. Given a suspect/POI from the case/investigation, the user (investigator) of the tool can obtain the sentiment expressed by him/her, narrow down to the document/evidence, and obtain the timelines. The tool is programmed to retrieve sentiment results and other data from the database for the investigator. The sentiments and data normalization can also be re-run/executed on demand from the tool.

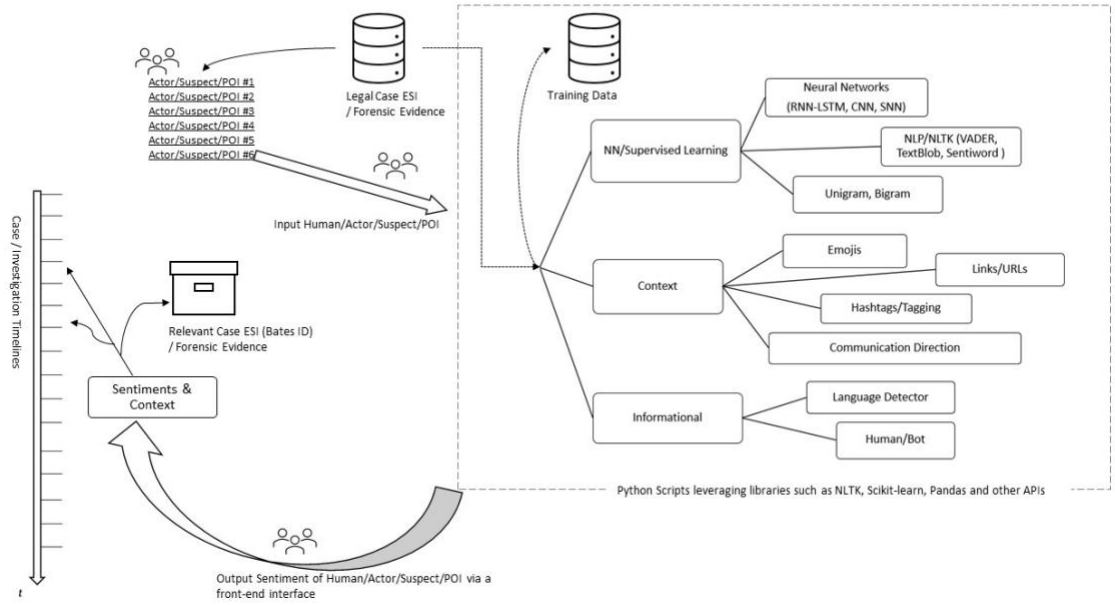

Fig. 3. Sentiment Analysis of a Suspect or Person Of Interest (POI) within the forensic investigation timeline or Legal Case eDiscovery scope

Dataset - Preparation and Normalization

The ESI datasets for each case were first assembled as flat-files from a variety of public sources. These flat files were then ingested using custom C#.NET programs into three SQL Server databases running on a local SQL Server instance. A .pst parser "Pstxy" [105] was used to parse emails. SQL tables were created for each type of data being injected. For training data, a separate database was created on the same instance. For each flat-file being ingested, file metadata was also identified and uploaded into the SQL tables. A new key column for bates-id (document id) was introduced to represent each ingested file. This resulted in three databases housing three different case ESI or digital forensic evidence for investigation. For each case ESI (databases), names of people (from the flat files) were randomly changed for fictitious actors/suspects of the case. Timestamp was randomly updated to reflect case/investigation timelines. Data for retweets, case timelines, followers, likes, the direction of communication and reaction was randomly added to each

post/tweet. Care was taken to randomize at every opportunity to avoid bias. For each type of case/investigation ESI (evidence) that is now on database tables, text data was normalized into separate tables. Sentence identification methods were employed to parse large documents and paragraphs. Each sentence was stored as a row in the normalized tables. Emojis, abbreviations, glyphs, grapheme clusters, email addresses, social media identifiers or handles, hashtags, acronyms, and URLs were extracted into separate columns and stored in the normalized tables. This was done as input text for analytics can sometimes consist of garbage/Html text, especially when dealing with Unicocde (UTF-8 encoded) and umlauts. Also, such information, when extracted, can help build a context for the investigator or legal mind using the proposed analytics software discussed in Chapter VII.

Sentiment Analysis using Supervised and Hybrid Learning

IMDB movie reviews sentiment dataset [104] was used for supervised learning. Feature selection process was not undertaken for analytics as the only feature to focus on was the text from the normalized tables. Different approaches outlined below were taken for sentiment analysis using Python. Database connection was established to first query the data from the database tables to train the model before applying it against the normalized data. The reason to use different approaches was to provide the investigator with choices and seek feedback on correct/incorrect sentiment categorization, thereby update the training data in a loop. The below NLP/NLTK and Neural Networks techniques were used. Randomization during labeled data selection for training and testing was avoided to allow for uniformity in results for the user of the custom tool, lest each run of analytical algorithms against case evidence would end with confusing different results confusing the user. The various approaches used are listed below.

1. Applied VADER (Valence Aware Dictionary and Sentiment Reasoner) [106] to each of the normalized case datasets in a supervised learning model using training and testing data in a 80-20 ratio (20000 for training and 5000 for testing) with no randomization. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to feelings and sentiments expressed in social media by considering individual tokens for sentiment analysis. VADER was installed using the command “pip install vaderSentiment” at the terminal window. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a separate database table. The SentimentIntensityAnalyzer class methods provided a sentiment intensity score to each text sentence.
2. Applied SentiWordNet [107] (an opinion lexicon derived from the WordNet [108] database) against each of the normalized case datasets in a supervised learning model using training and testing data in a 80-20 ratio (20000 top samples for training and 5000 bottom samples for testing) with no randomization. SentiWordNet is a lexical resource for opinion mining and is publicly available for research purposes. Python’s NLTK provides both SentiWordNet and wordnet classes for import. SentiWordNet approach computes the polarity of the words and averages the value. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each of the datasets.
3. Applied TextBlob [109] against each of the normalized case datasets in a supervised learning model using training and testing data in a 80-20 ratio (20000 top samples for training and 5000 bottom samples for testing) with no randomization. TextBlob is a Python (2 and 3) library for processing textual data.

TextBlob is a python library offering a simple API to access its methods and perform basic NLP tasks. Naïve Bayes was used as the classifier from Textblob library of classifiers as it offered better accuracy compared to MaxEntClassifier. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each dataset.

4. Applied Unigram approach along with Python's NLTK SentimentAnalyzer against each of the normalized case datasets in a supervised learning model using training and testing data. Unigrams or 1-gram is an N-gram with simply one string in a text. A Naïve Bayes classifier was used. Due to memory limitations, the training set was limited to top 3000 samples, and the testing set was bottom 600 samples (out of a total of 25,000 in the labeled dataset) with no randomization. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each dataset.
5. Applied Bigram approach against the normalized case data in a supervised learning model using training and testing data. Bigram or 2-gram is an N-gram that is typically a combination of two strings or words that appear in a text. A Naïve Bayes classifier was used. Due to memory limitations, the training set was limited to top 5000 samples, and the testing set was bottom 1000 samples (out of a total of 25,000 in the labeled dataset). Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each dataset.
6. Applied a Long short-term memory (LSTM) (Recurrent Neural Network) against each of the normalized case datasets in a supervised learning model using training and testing data. Trained on 16000 samples and validated on 4000 samples from

the labeled dataset with 6 Epochs, 128 neurons and no randomization. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each dataset.

7. Applied a Convolutional Neural Network (CNN) against each of the normalized case datasets in a supervised learning model, using training and testing data. Trained the model on 16000 samples and validated on 4000 samples from the labeled dataset. Used 6 epochs using GloVe embeddings [110] to create our feature matrix, one dimensional Convolutional layer with 128 features, kernel size of 5 and activation function as relu. Finally, a dense layer was added and used activation function as Sigmoid. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each dataset.
8. Applied a Simple Neural Network against each of the normalized case datasets in a supervised sequential learning model using 16000 samples for training and 4000 samples as testing data, embedding layer of 100, final dense layer with activation function as Sigmoid, and 10 epochs. Used using GloVe embeddings to create our feature matrix. Sentiment polarity against each sentence and accuracy of the model was calculated and stored in a database table for each dataset.

Presentation

A custom Windows Forms application screen was designed and developed to help steer the case investigator to use the sentiment analysis of a case suspect. MS-SQL stored procedures and queries were written to automate the display and enrich context by leveraging data about URLs, emojis, timelines, etc. A simple logic was added to determine if the suspect was a bot instead of a human.

Analysis

Upon application of sentiment analysis algorithms, and upload of the case dataset, the investigator can now pick a suspect/POI from the case to obtain his/her sentiments as classified by the different models. Each analytical model yielded sentiment results and accuracy based on size of training and test sets. ROC and AUC were calculated for cross verification.

1. VADER Using Python's VADER library, the whole labeled dataset was used in an 80-20 ratio for supervised learning (top 20,000 samples for training and bottom 5000 for testing). A model accuracy of 69.1% was achieved. Fig. 4 displays the model's ROC curve with AUC = 0.69.
2. Using Sentiwordnet form Python's NLTK library, the whole labeled dataset was used in a 80-20 ratio for supervised learning (top 20,000 samples for training and bottom 5000 for testing). A model accuracy of 66.7% was achieved. Fig. 5 displays the model's ROC curve with AUC = 0.67.
3. Using Textblob approach and Naïve Bayes Classifier, top 2000 samples were used for training and bottom 400 for testing due to memory limitations. A classifier accuracy of 84.3% was achieved. Fig. 6 displays the model's ROC curve with AUC = 0.68.

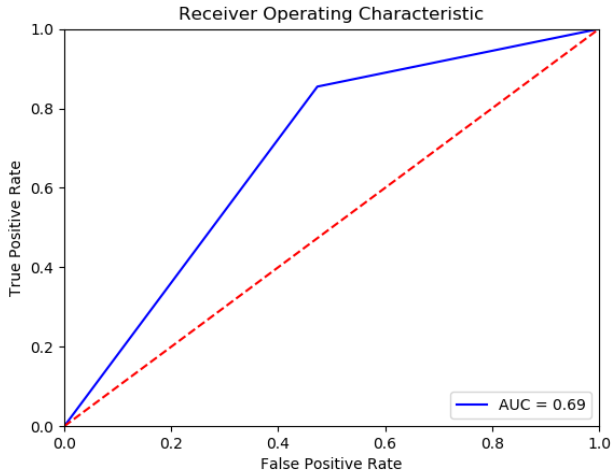

Fig 4. ROC using VADER (Valence Aware Dictionary and Sentiment Reasoner)

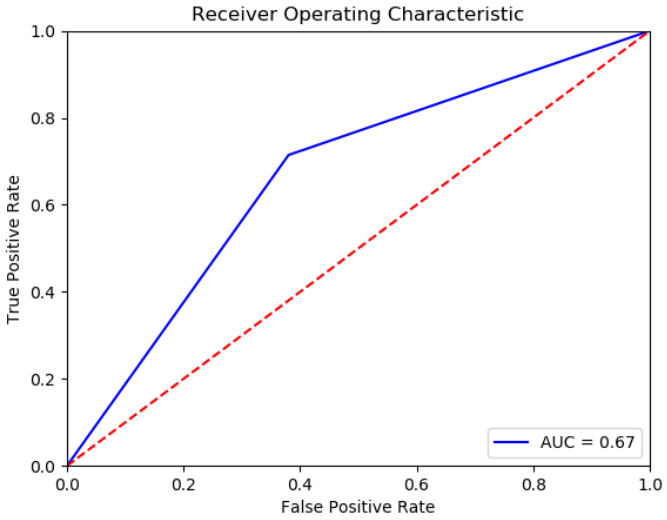

Fig 5. ROC using SentiWordNet

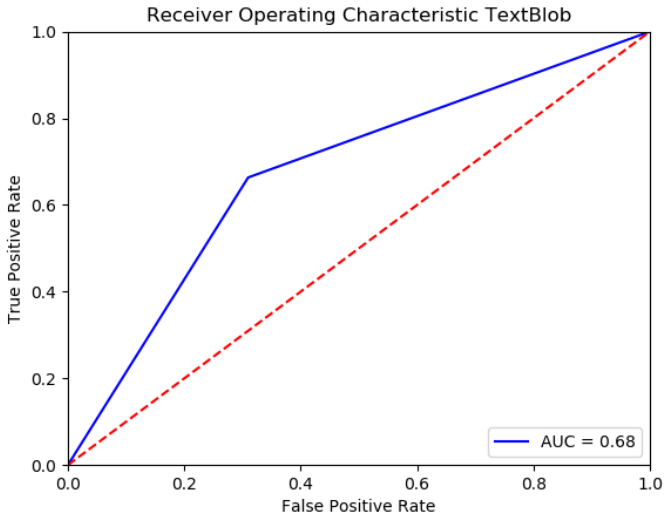

Fig 6. ROC using TextBlob

4. Using Unigram approach, NLTK SentimentAnalyzer, and Naïve Bayes Classifier, a model accuracy of 78.2% was achieved. Due to memory and CPU limitations, small sizes of training and test labeled data was used. Also, increasing the training set size lowered the accuracy. Fig. 7 displays the ROC curve with AUC = 0.5. However, the Precision is high (0.72), Recall is high (0.85) and F-measure is high (0.78). High accuracy but a low AUC value implies that the training features may be imbalanced i.e., there are much more negative sentiments than positives taken in consideration during training. The apparent discrepancy has to do with the lack models' success at identifying true negatives.
5. With the Bigram approach increasing training set size decreased the accuracy and caused memory and high CPU usage issues. Thus, a subset of the labeled dataset considered for training, and model accuracy achieved was 54.6%. Fig. 8 displays

the ROC curve with AUC at 0.5. However, the Precision was 0.54, recall was low at 0.02 and F-measure was 0.042.

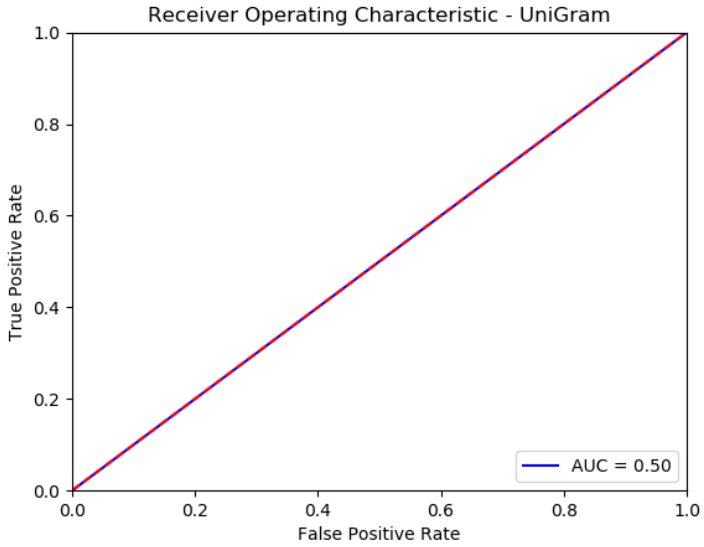

Fig. 7. ROC using Unigram

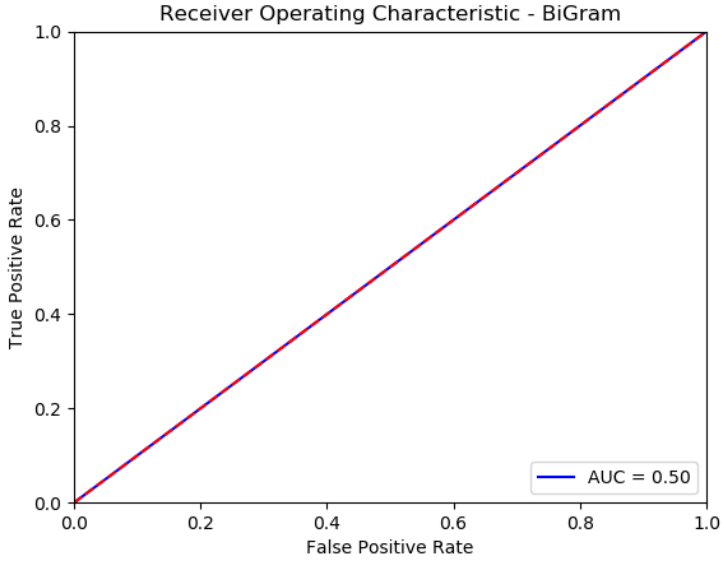

Fig. 8. ROC using Bigram

6. Using the RNN approach, to compile the model and used the adam optimizer, binary cross entropy as our loss function and accuracy as metrics. Next a sequential model was initialized followed by the creation of the embedding layer. Next, a LSTM layer with 128 neurons was created. GloVe embeddings was used to create the feature matrix. The model's accuracy was 80.6%. Fig 9 displays the ROC curve with AUC at 0.92.

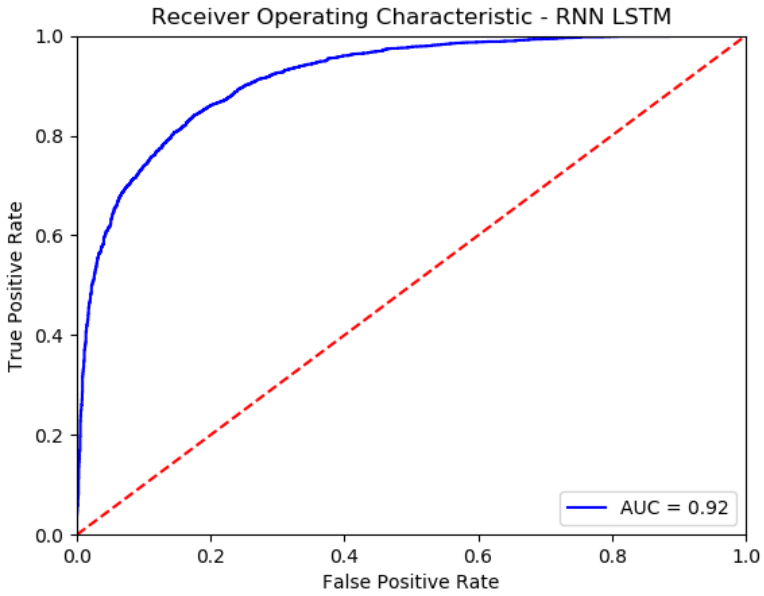

Fig. 9. ROC using Recurrent Neural Network (RNN) with LSTM

7. Using the CNN approach, to compile our model, the adam optimizer was used along with binary cross entropy as our loss function and accuracy as metrics. We used GloVe embeddings to create our feature matrix. The model's accuracy was 84.1%. Fig. 10 displays the ROC curve with AUC at 0.92.
8. Using the Simple Neural Network approach, to compile our model, the adam optimizer was used along with binary cross entropy as the loss function and

accuracy as metrics. We used GloVe embeddings to create our feature matrix. The model's accuracy was 69.4%. Fig. 11 displays the ROC curve with AUC at 0.76.

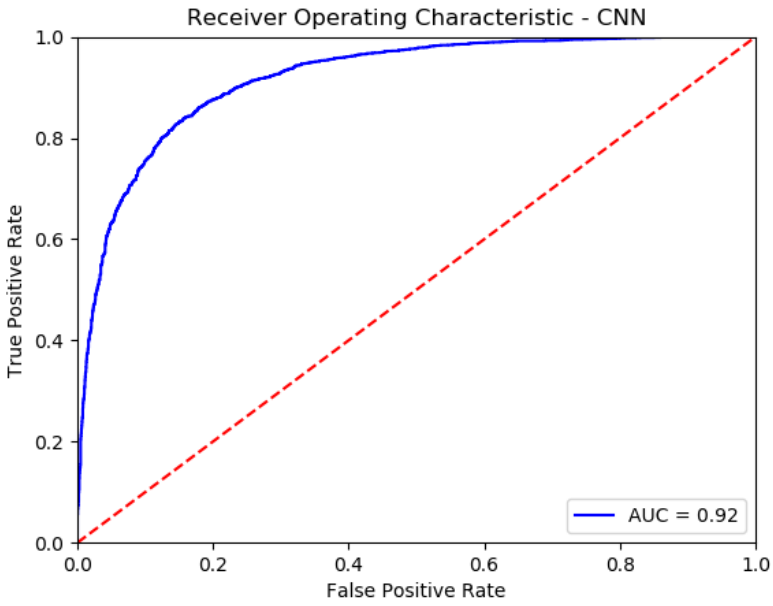

Fig. 10. ROC using Convolutional Neural Network (CNN)

Table 4 summarizes the results all the models/approaches used in our experiment.

Sentiment analysis for natural language processing (NLP) applications has become easy and widely accessible thanks to the rising popularity of Python and open-source NLP programs like TextBlob and VADER. However, these packages have some drawbacks. Although NLP tools like TextBlob and VADER are excellent, they are not very accurate for tasks requiring sentence-level sentiment classification [111]. The reason behind their weaknesses is due to the fact that they are based on the bag-of-n-grams model. The bag-of-n-grams model treats natural language as a collection of n-grams, as the name suggests. As all the words in a phrase are tossed into a mixed bag of words, the bag-of-n-grams

model does not consider the word order of natural language. However, the semantics of natural language are strongly related to word sequence order. Neural network-based models like CNN, LSTM, DNN, attention and Simple Neural network take the word sequence information into modeling consideration by design and tend to perform well over the n-grams approach. TextBlob and VADER are still great prototyping tools and fantastically easy to use, but, can fall short in accuracy when compared to neural network based models. The SentiWordNet (SWN) model is based on a word dictionary, which builds on top of the original Princeton WordNet dictionary by adding sentiment scores (positivity/negativity) to each word and an objectivity score. The scores all add up to 1 and are split between positivity/negativity and objectivity. The SentiWordNet approach again does not take into account the semantics of natural language that are strongly related to word sequence order.

The AUC score helps us quantify the model's ability to separate the classes by capturing the count of positive predictions which are correct against the count of positive predictions that are incorrect at different thresholds. By analogy, the higher the AUC, the better the model is at distinguishing between sentences with positive and negative sentiments. We can deduce from Table 4 that the Neural Network based models (CNN, LSTM and Simple Neural network) have a high AUC compared with (SentiWordNet, TextBlob and VADER). Since the Unigram and Bigram experiments were not performed using the full training set (due to computational resource limitations), they are not considered in these comparisons, but since they belong to the n-gram type of models, they may still fare lower in AUC scores than the neural network-based models.

A custom Windows Forms application as shown in Fig. 12 was developed to help steer the case investigator to use the sentiment analysis against case suspects/POI. At run-time, all suspects/POI in the case dataset (ESI) was listed on the screen. Upon the user choosing one of the case suspects/POI, sentiments from various algorithms listed above were displayed along with accuracy. Database stored procedures and queries were executed by the custom application to automate the display of information on the screen. A logic for bot detection for tweet/post was also incorporated on the tool. To detect bot activity, few indicators like - high volumes of activity, a high percentage of retweets, many followers with less followed, etc. were programmed. Average sentiment on sentences per document (bates id) per person was then displayed per algorithm. Detailed sentiments of the person *per* document was also presented on the screen. Other details such as timelines and context were also displayed. This would be helpful to the case investigator as background details of the suspect/POI. The tool also allowed for the case investigator to manually mark a sentiment as correct or incorrect and this was fed back into the labeled training dataset to be reused in the next analysis run along with the labeled dataset. The tool also allowed for the case investigator to pick specific date range within the case timelines as a filter.

TABLE 4
Summary of results from various sentiment analysis algorithms
(* partial labeled dataset used for supervised learning)

Model/Approach	AUC	Accuracy (%)
CNN	0.92	81.4
LSTM (RNN)	0.92	80.6
SentiWordNet	0.67	66.7
Simple Neural Network	0.76	69.4
Unigram*	0.50	78.2
Bigram*	0.50	54.6
TextBlob	0.68	84.3
VADER	0.69	69.1

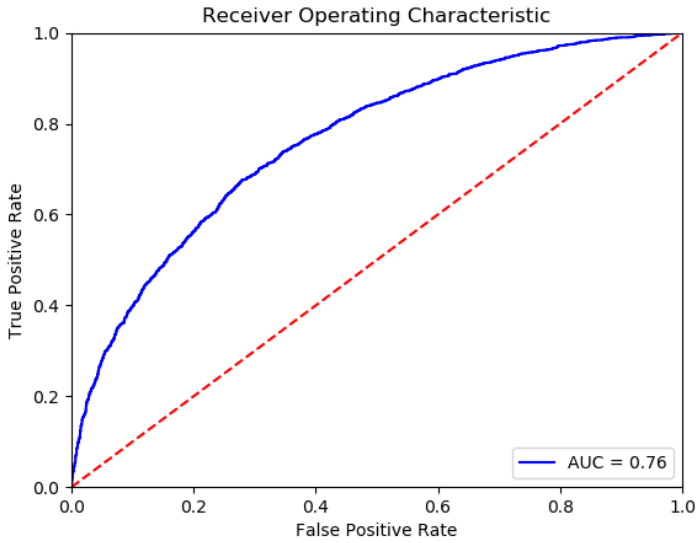

Fig. 11. ROC using Simple Neural Network

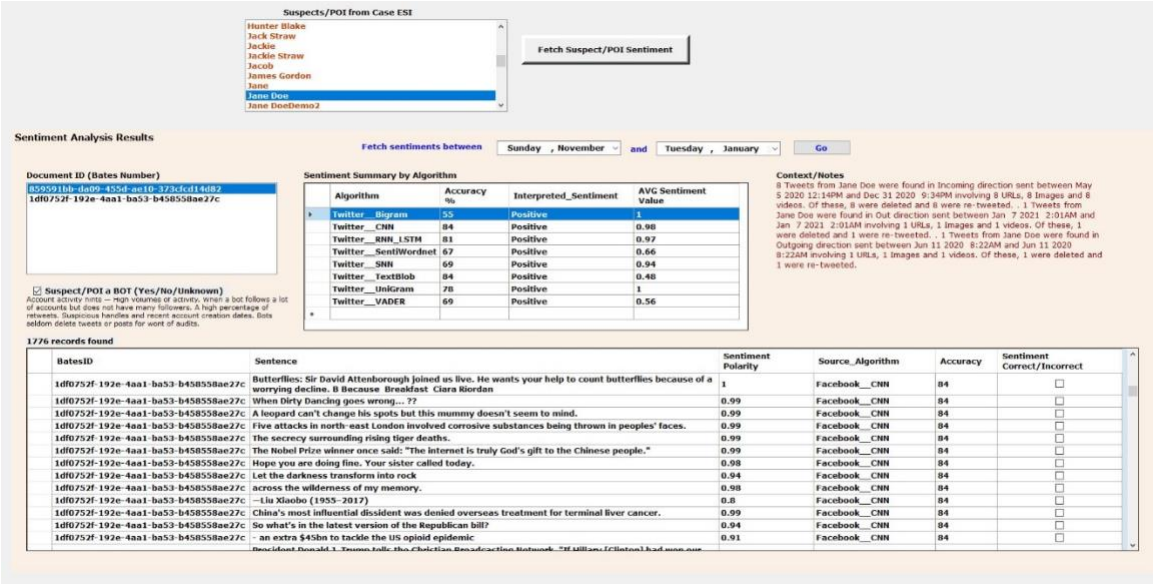

Fig. 12. Custom software developed for sentiment analysis

Summary

As electronic case artifacts and their accompanying data have expanded in volume and originate from a variety of sources, using machine learning in legal analytics and digital forensics to speed-up the investigation can be quite beneficial. A typical case to investigate may involve processing large amounts of electronic data in the quest for something such as sentiments expressed by suspects involved in the case. There may be a few suspects in the case that the investigator may want to focus on and manually analyzing evidence to build the sentiments expressed by the suspect can be time consuming. This research demonstrates the use of various machine learning and neural network approaches to process legal/forensic case evidence (ESI) and mine sentiments of suspects involved in the case. Fictitious (synthetic) case datasets were assembled from custom and public sources, and various analytical approaches for sentiments coupled with a custom software was developed. In addition to displaying a comparative viewpoint, the use of multiple analytical approaches allows the investigator to pick a particular approach over the other and purse their investigation. This avoids bias in analytical technique selection from the very beginning. The custom software allows for fine-tuning the training dataset over time due to a user feedback loop, thereby allowing for improved model accuracy over time and use. Thus, the software helps reduce analysis time, reduces costs of the case investigator to analyze electronic data from the case pile for suspect sentiments and reduces rework effort. Data cleansing techniques (preprocessing) employed on case ESI, and the quality of the training dataset used can greatly affect the overall results of the various analytical models. Overall, this proposed approach gives insights into suspects of the case to retain or eliminate them during an investigation.

CHAPTER V

FINANCIAL FRAUD DETECTION OF CASE SUSPECT

Evidence for a real-life forensic investigation of financial fraud was hard to find in public for academic research. Thus, the authors felt the need to customize and build random fictitious electronic evidence (ESI) for this experiment [112]. The experiment was carried out using an Intel(R) Core (TM) i5-3470 CPU @ 3.20GHz 16 GB RAM PC and a 64-bit Windows 10 operating system. Software used was Python, SQL Server 2019, and Visual Studio 2019. This experiment is solely to showcase an umbrella approach to tackling securities/financial fraud investigations. To avoid bias, the experiment results are published as-is, and no attempt was made to withhold wayward results or showcase only high-fidelity results.

Insider Trading

An insider's motive (intent) to buy or sell stock using privileged information unknown to the public is key to indicators of insider trading fraud. The logic for detecting insider trading considers four different factors as shown in Fig. 4. The essential aspects of this logic are the provision for estimating an individual's risk, as well as static and temporal abnormalities combined with machine learning approaches. The crux of the logic is the intent exhibited by a suspect when trading stock. Intent can be gathered from the communications of the suspect and then correlated against stock prices of the same timestamp. The key to establishing insider trading is the suspect's knowledge of privileged information. This is indicated by the suspect's attendance at crucial meetings, access to IT systems, and access to coworkers who may have access to this information. A high-risk employee profile is defined in terms of Financial IT systems privileged access, action

owner of past audits, etc. Sentiments of a suspect's communication can greatly assist with legal arguments and thus sentiments across case evidence were highlighted for the investigators. Fig. 6 shows our proposed software implementing this logic and handling insider trading scenario across whole evidence.

Pump & Dump

Since we are determining intent from textual evidentiary data, a simple logic for pump and dump (P&D) can be implemented using suspect's intents. If we observe a pattern of intent to "buy" stock followed by an intent to "sell" and this pattern correlates to stock price increase followed by a drastic fall, then we can conclude there is an indication of P&D. For P&D the suspect need not be an employee of the company and thus any such metadata collected by the tool was ignored. Fig. 13 describes this logic. This logic can be further tuned for parameters such as the amount of price increase (pump), the amount in price decrease (dump), the volume of stock sold between the timestamps of this pattern and communication time gaps (days, hours) to trigger a P&D indicator. For the sake of simplicity, the logic demanded a minimum of three continuous text communication evidences of stock purchase intent, followed by a single textual communication evidence of stock sale intent. This should then correlate to an increase in stock price (due to buy/pump intent) and volume followed by an immediate decrease (due to sell/dump intent). Fig. 17 shows our proposed software implementing this logic and handling P&D scenario across whole evidence.

Experiment Design

The aim of this experiment is to propose an approach that helps the investigators to investigate financial fraud, especially cases of insider trading and Pump & Dump schemes.

Case evidence was mined for human intent, unlabeled data was labeled using unsupervised learning, and various algorithms were implemented along with risk ranking, suspect profiling, and sentiment analysis to arrive at fraud indicators. To arrive at such conclusions, a mix of direct mining of evidence coupled with machine learned predictions from labeled data is employed. Fig. 14 shows a high-level investigation approach with suspect profile,

BERT Intent	Date	Stock price
Sell	10/12/2020	18.31
Buy	10/13/2020	19.22
Sell	10/14/2020	20.01
Sell	10/15/2020	19.33
Buy	10/16/2020	20.72
Buy	10/17/2020	21.22
Buy	10/18/2020	22.01
Buy	10/19/2020	27.36
Sell	10/20/2020	18.28

Normal Trading

P&D

Fig. 13. Pump and Dump (P&D) logic using Intent

suspect intention, stock value, and risk as inputs producing a Boolean indicator of fraud as output along with the source of evidence. For ease of understanding, the proposed fraud detection approach utilizes the three buzzwords of any investigation, namely: means, opportunity and motive. While this detection approach can address means and opportunities to a certain degree, it is left to the investigators and prosecutors to establish a motive. However, for ease of understating, the motive is taken as profiting from stock prices. Fig. 15 highlights the various machine learning and automation methods/techniques leveraged under the proposed detection umbrella. The reason for proposing multiple approaches/ methodologies is that investigators are not bound by one but instead have a

mix of approaches to choose from. Also, each analytical approach has a built-in user feedback feature that, when triggered by the investigator, will contribute back as user-labeled data that can be reused for supervised learning.

Dataset Preparation

The datasets used for this experiment was from prior research [112]. The key types of data were from fictitious emails, Facebook posts, Tweets, WhatsApp/SMS messages, and random MSWord documents. Data was stored in SQL tables identified by their source/document identifier known as bates number/ID. Each email and MS Word documents were further broken down into sentences and stored in a separate SQL table. Data needs to be processed for analytics as there can be occurrences of emojis, hyperlinks, stop words, etc. that can inhibit the analytical process [113]. All textual data was pre-processed using Natural Language Processing (NLP) techniques such as tokenization, stop words, stemming and lemmatization. All suspect names, key event dates, textual data and stock symbols used are solely for demonstration purposes and bear no resemblance in any shape or form in real life.

1) Reddit Data: Stock trading and finance-related discussion data from Reddit forums was collected via allowed Reddit APIs [114]. For simplicity, subreddits (community/channel/forum) considered were Wallstreetbets and Investing. Python scripts were written and executed between Nov/06/2021 and Nov/14/2021 to read each subreddit data and write into .csv files that were later stored as SQL tables. A total of 155,651 rows of Reddit data was collected. This step can be altered if the investigation team has quality labeled stock-trading data for supervised learning.

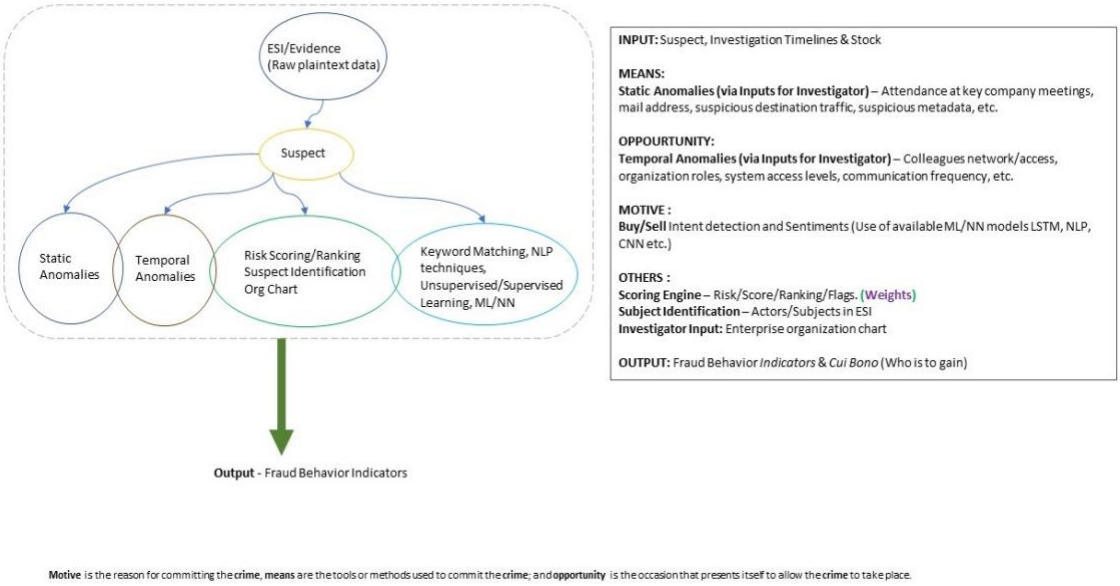

Fig. 14. Financial fraud detection – High level approach

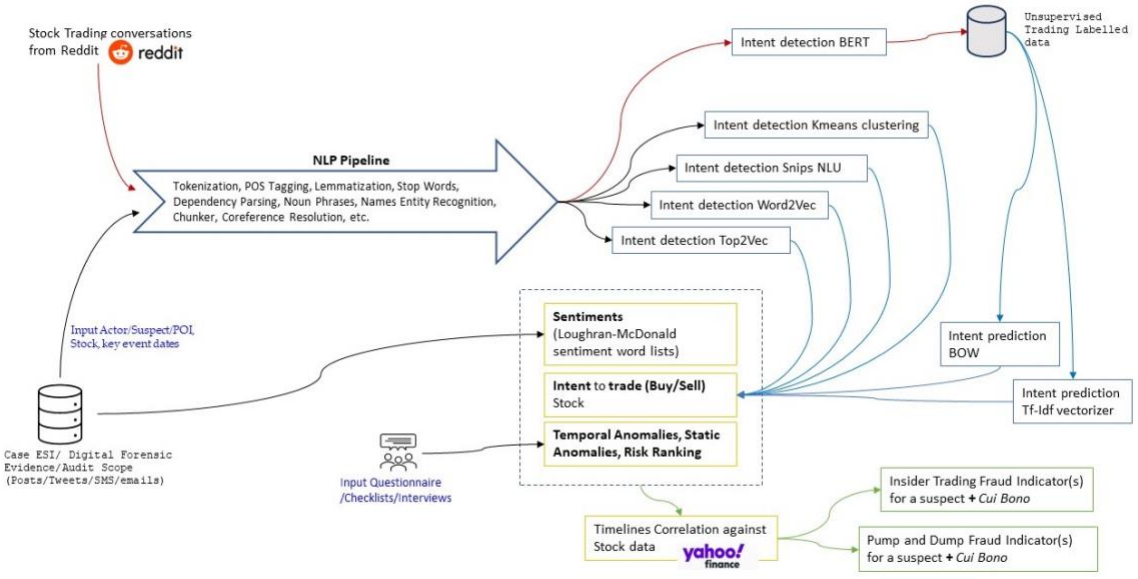

Fig. 15. Financial fraud detection process involving various approaches

2) Yahoo Finance Data: Historical market data from Yahoo Finance was obtained as needed using the python module yfinance [115]. The module yfinance is a python

module that uses Yahoo! Finance's API and returns stock, cryptocurrency, forex, mutual fund, commodity futures, ETF, and U.S. Treasury financial data. Python scripts were written and executed via C#.NET on the prototype tool. These scripts also inserted data into SQL tables when executed.

3) Ancillary Data: For various automation steps, ancillary data such as stock ticker/symbol data, emojis, emoticons, stop words, etc., were assembled from the Internet. Few data files were stored in SQL databases, while the rest were stored as flat files. All stock data was limited to NASDAQ, NYSE, and NYSE stock exchanges that can be further expanded to other exchanges.

BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for NLP developed by Google [116]. BERT can be used in a wide variety of NLP tasks such as question answering (SQuAD v1.1), Natural Language Inference (MNLI), and others. A python script was written to perform a sort of unsupervised classification of textual Reddit data into buy, sell, or other based on similarity. This approach helps us label the Reddit data in an unsupervised way. After text data preprocessing, creation of target clusters using Word2vec and gensim was performed, followed by word Embedding with transformers and BERT. The gensim package has a function that returns the most similar words for any given word. Lastly, observations to clusters were assigned by their cosine similarity and model's performance was evaluated. Classification results were stored in SQL tables as labeled data using BERT.

1) TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) statistical approach determines how important a word is by weighing its frequency of occurrence within

the document. After data preprocessing (word cleaning, stop words removal, hyperlinks, stemming, lemmatization), the data from BERT was split into training & testing subsets. A Naive Bayes classifier was used to fit the training data, and predictions were obtained with the test dataset. Model's accuracy, precision, recall, confusion matrix, and ROC was obtained. This model was then applied against text from each document (identified by bates number) from the investigation caseload and prediction results of buy/sell/other were stored in a SQL table.

2) BOW: The Bag-of-Words (BOW) model builds a vocabulary from a corpus of documents and counts how many times the words appear in each document. A python script was created for implementing BOW. After data preprocessing (word cleaning, removal of stop words, hyperlinks, stemming/ lemmatization), the labeled dataset (using BERT technique) was split into training & testing subsets. The TF-Idf vectorizer and Naive Bayes classifier was applied to transform and predict test data. Model's accuracy, precision, recall, confusion matrix and ROC was obtained. This model was then applied against text from each document (identified by bates number) from the investigation caseload and prediction results of buy/sell/other were stored in a SQL table.

K-Means

This approach involves unsupervised text clustering using NLP and K-Means. Against the Reddit data, the TF-Idf vectorizer was applied using a python script followed by clustering using K-Means to find top 3 clusters. After data preprocessing (word cleaning, removal of stop words, hyperlinks, stemming/ lemmatization), the clustering

model was then applied directly against each document (identified by bates number) from the investigation caseload and similarity results were stored in a SQL table.

Top2Vec

Top2Vec [117] is an algorithm for topic modeling and semantic search in a large collection of documents. Top2Vec utilizes Doc2vec to first generate a semantic space that consists of word and document vectors in a continuous representation of topics. There was no need to remove stop words as such words will appear in almost all documents present in the corpus, therefore being equidistant from all topics. They will not appear as a nearest word to any specific topic. Stemming/lemmatization was not implemented, but text was cleaned for punctuation and made lowercase. A python script was created to implement Top2Vec against case evidence for keywords “buy” and “sell” and similar semantic words. Results were stored in a SQL table.

Word2Vec

Word2vec is a popular technique to learn word embeddings using deep learning and a two-layer neural network. Its input is a text corpus, and its output is a set of vectors wherein semantically similar words are placed close to each other. Word2Vec model comes in two flavors: Skip Gram Model and Continuous Bag of Words Model (CBOW). A python script was created using gensim library implementing both Skip Gram Model and CBOW approach of Word2vec directly against the case evidence. The script computed the similarity of words to “buy” and “sell” in each bates number of the investigation evidence caseload. Results were stored in a SQL table.

Snips

Snips NLU [118] is an open-source Natural Language Understanding (NLU) python library that allows parsing sentences written in natural language and extract structured information [119]. The NLU engine first detects the from text the intention of the user, then extracts the parameters (called slots) of the query. As required by Snips, a json/YAML file was fitted in the SnipsNLUEngine with custom utterances of buy/sell intent and stock symbols. Fig. 16 shows a snapshot of this file contents. A good alternative to Snips NLU was Rasa NLU [120]. However, Snips NLU has been proven to be better than Rasa NLU [121], [122] and thus used in this experiment. Snips results were stored in SQL tables.

Sentiment Word List

Suspects can display sentiments that can help in profiling. The Loughran-McDonald sentiment word lists [123] was used to perform sentiment analysis as this was specifically built and is maintained for textual analysis related to finance. This added information could help investigators better understand the behavioral aspect of the suspect at a particular timeline corresponding to the text origin. A python script was created to implement the Loughran-McDonald sentiment word lists against each bates number and suspect in the case evidence pile. Results were stored in a SQL table.

Calendar of Key Events

To correlate key dates of events against specific evidence (example tweet or SMS sent date) and lookup of historical stock prices, a C#.NET input screen/form was created to ingest multiple dates and event details. This information was manually input by the

investigators and stored in a SQL table. For the sake of simplicity, only dates were considered although this can be extended to include the time of day.

Suspect Metadata

A C#.NET input screen/module was created to ingest suspect metadata such as their company designation, attendance at meetings (dates), and work hours. Such data can help support the investigation findings. This information was manually input by the investigators for suspects and stored in a SQL table.

Risk Profile and Ranking

A C#.NET input screen/module was created to ingest suspect risk metadata such as involvement in financial audits, access to key colleagues, finance systems access levels, elevated privileges if any, prior red flags from Human Resources (HR) department of the company and a prior victim of phishing. Such data can help build a risk profile, provide circumstantial/ observational evidence, and provide valuable insight of the suspect to the investigators. Risk ranking was based on a weighted approach of this metadata that can be customized by the investigators. The investigators manually input this information for suspects and stored in a SQL table.

Presentation

A Windows Forms (client/server) module was created using C#.NET as a prototype software [112] for use by the investigators. Fig. 17 and Fig. 18 show the main screens of this software when used for “insider trading fraud” and “pump and dump” detection. For an investigation to commence, the user first inputs suspect metadata, suspect risk details, and key event dates that lie within the scope of the investigation. The next step involves choosing the suspect, the company stock, and dates of interest. The user can then choose

```yaml
#Buy Intent during Stock Training
type: intent
name: Buy
slots:
  - name: ticker_buy
    entity: ticker
  - name: share_volume
    entity: vol
  - name: date
    entity: snips/datetime
utterances:
  - find me price of [ticker_buy](AAPL) for bulk order of [share_volume]
  - I need to add [ticker_buy](GOOG) to my portfolio
  - show me gains for [ticker_buy](DAL) [date](this evening)
  - for [date], will [ticker_buy](ABT) be good for buying?
  - would you recommend [ticker_buy](AXP)?
  - is this [ticker_buy](APA) worth buying
  # few synonyms of buy
  - purchase
  - procure
  - acquire
  - obtain
  - pick up
  - acquiring
  - acquisition
  - interested
  - get involved in
  - promising
  - position # I would suggest we position a block of 2,000 shares.
  - gain
  - profit
  - benefit
  - investment
  - buying
  - profit
---
# Tickers Entity
type: entity
name: ticker
automatically_extensible: true # default value is true
use_synonyms: true # default value is true
matching_strictness: 0.9 # default value is 1.0
values:
  - [OEDV , Osage Exploration and Development Inc.]
  - [AAPL , Apple Inc.]
```

Fig. 16. Snips json/YAML logic

the option to find evidence of insider trading or evidence of a pump and dump. In the case of pump and dump scenario, certain data elements of the suspect collected earlier may not be relevant such as job designation, system access levels, or association to the company.

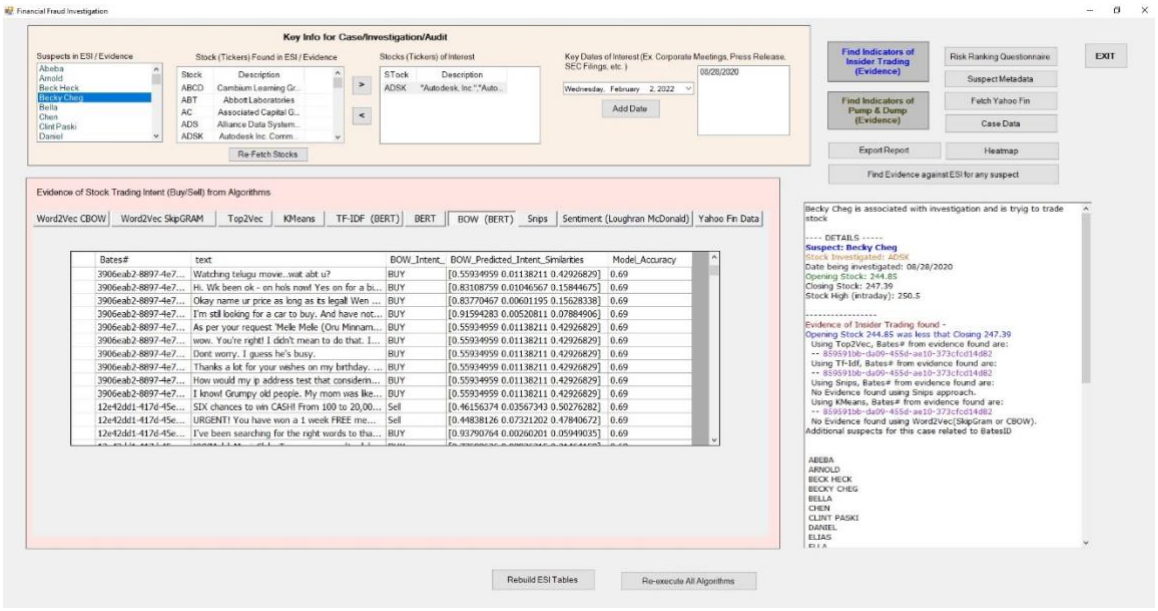

Fig. 17. Screen of custom software for use by case investigator for Insider trading. [Note: Trade stocks and names shown are purely for academic study and have no bearing on an event/person/investigation.]

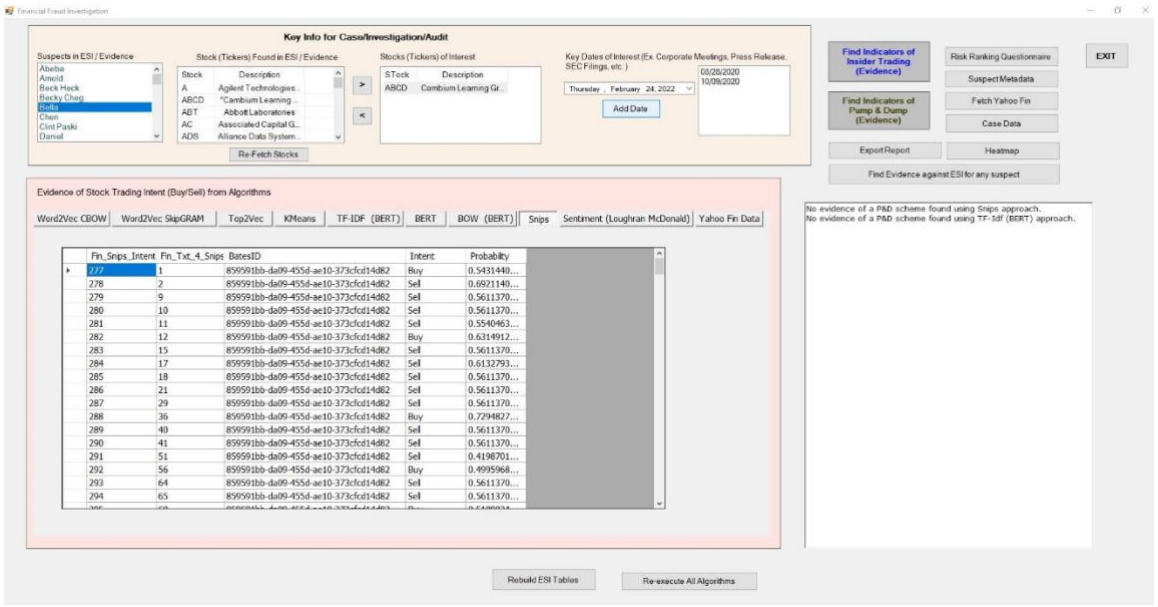

Fig. 18. Screen of custom software for use by case investigator for Pump and Dump (P&D) scheme. [Note: Trade stocks and names shown are purely for academic study and have no bearing on an event/person/investigation.]

On the prototype software, upon user action to find evidence of insider trading or pump & dump, data stored on various SQL tables is correlated against historical stock data obtained from Yahoo Finance API. Results of correlation is then displayed on the screen along with

sentiment data and risk ranking for the suspect, pointing eventually to the bates number if there as any evidence found. If no evidence was found, a suitable message was displayed on the screen. The user screen allows for the download of a report and re-run any of the abovementioned algorithms. The software also allows for storing other investigation related details.

Analysis

This research combines supervised learning and unsupervised learning to help locate fraud indicators in a stack of electronic evidence. This approach is known as Hybrid (supervised/unsupervised) learning. The algorithms used in this research can vary and can be improved with user feedback (fine tuning). This approach is suitable for an investigation team that has no prior labeled data on trading intents. They can start with unlabeled data and over the period of many investigations, build a quality dataset.

1) Quality of Case Data: Typical case data can be a collection of files on electronic/computer systems housing any information related to the scope of investigations. Often the initial collection volume is more than required as the investigation scope may not be well defined and kept broad. The eDiscovery EDRM [7] model can be applied to the process of vetting and analysis to filter out the irrelevant data and retain data as critical evidence for legal arguments. The process of vetting native format involves a ton of data processing tasks such as data masking, redaction, culling, etc. A major problem that investigators may encounter is the language used in evidence. When Word2Vec was used against a good dataset like glove-wiki-gigaword-50, similar words were found to be more accurate. However, our case evidence data for Word2Vec was not as good of linguistic

quality as that of glove-wiki-gigaword-50 and not encompassing English language. Thereby our results were not as expected. This is a common real-life scenario that investigators will encounter as communication data in case evidence these days is not the standard English language. This holds good to other prominent world languages as well as the Internet has been bemoaned as the downfall of the written word, pronunciations, and grammar. For example, multiple languages can be found mixed with each other in a single SMS text. Thus, quality of case data can vary from case to case and care must be taken to first analyze linguistics within the textual evidence followed by management of non-textual data such as media, biological evidence, etc.

2) Labeled/Unlabeled stock trading data: Due to the unavailability of a public labeled dataset on stock trading intents, case investigators may need to create a labeled dataset for supervised learning. The proposed approach can be executed against any historical case evidence data to arrive at a labeled dataset. Similarly, few public online sources of trading discussions such as news articles, discussion forums and financial market watch comments can be assembled to build an unlabeled dataset. The BERT approach discussed in this research can then be applied to label this dataset. A manual review of intents can then be completed to validate the quality of data upon which the labeled dataset on trading intents can then be used for analytical experiments. All text data must be carefully processed for hyperlinks, emojis, gifs, emoticons, smileys, abbreviations, etc. [32]. Such data should not be left behind but rather processed into their textual equivalent. This can be an uphill

task for the investigation team if there is no such labeled dataset to begin with, but once created, can be reused with periodic updates for many investigations.

3) Supervised/Unsupervised learning: The datasets used in the experiment were randomly picked and assembled to mimic typical case evidence and investigation. Twitter data, WhatsApp data, SMS data, emails, random custom MSWord documents, and Facebook data constitute the case evidence. Thus, the accuracy of models and results of the experiments were solely for demonstration of the approach. The accuracy of the TF-IDF model was found to be at 65%. The probability of snips in determining a "buy" or "sell" intent was between 0.38 and 0.79. The Top2vec algorithm had a 0.2 for similar word score. The similarity of words to "buy" or "sell" found using Word2Vec was between 0.97 and 0.99. The Fig. 19 displays the ROC curve of the BOW approach for "Buy", "Sell" and "Other". Fig. 20 displays the ROC curve of the TF-IDF approach for "Buy", "Sell", and "Other". BOW method was employed directly against the evidence while TF-IDF was employed against the labeled Reddit data (after using BERT to label this data). Thus, we cannot compare the BOW approach against TF-IDF as both are employed against different datasets. Case investigators can ignore or retain a model based on spot checks and manual analysis of evidence. This umbrella approach provides investigators with various approaches towards determining fraud indicators.

To summarize our experiment metrics, the BERT model achieved a 41% accuracy when predicting financial fraud intent of a suspect. The TF-IDF model prediction accuracy was

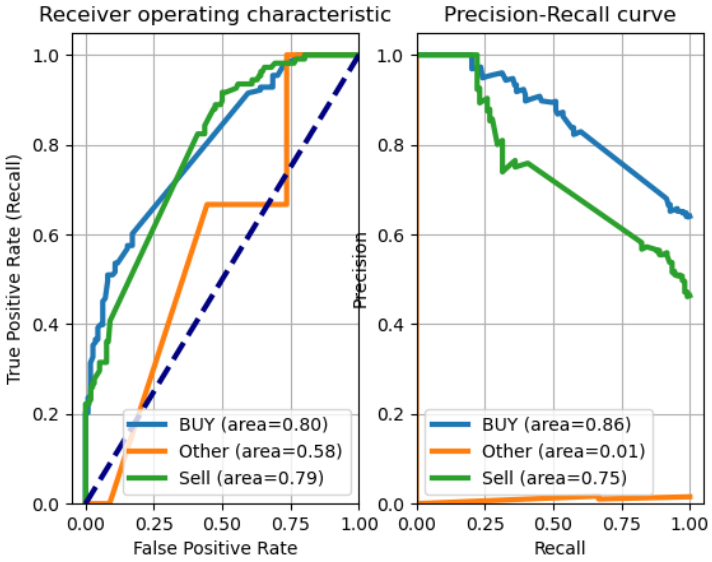

Fig. 19. BOW approach – ROC, precision, and recall

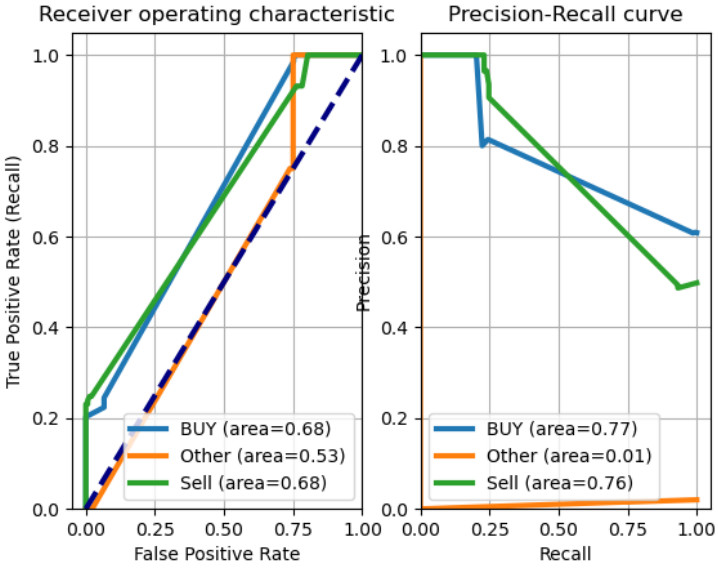

Fig. 20. TF-IDF approach - ROC, precision, and recall

65% and the BOW model prediction accuracy was 70%. The probability of Snips NLU in determining a “buy” intent was 79% and “sell” intent was 78%.

The BERT model by Google is based on the concept of transformers and considers the whole sentences for labelling instead of words. BERT is considered as the state-of-the-art language model for NLP and was specifically trained on Wikipedia (about 2.5 billion words) and Google's Books Corpus (about 800 million words). Masked Language Model (MLM) enables/enforces bidirectional learning from text by masking (hiding) a word in a sentence and forcing BERT to bi-directionally use the words on either side of the covered word to predict the masked word [124]. In our use of BERT for unsupervised labelling of sentences for intents, the BERT code can be further fine-tuned for better accuracy when defining dictionary of clusters. Fig. 21 displays the code used for defining the clusters in BERT logic. This fine-tuning comes at the cost of CPU usage and longer program execution time. Post labelling, the two models to predict indicators of financial fraud are Bag of Words (BOW) and TF-IDF (Term Frequency Inverse Document Frequency). BOW creates a set of vectors containing the count of word occurrences in the document/sentence. The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier. The TF-IDF model contains information on the more important words and the less important ones (rare) as well and gives larger values for less frequent words in the document corpus. To account for the fact that some words are used more frequently than others overall, the TF-IDF score increases according to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the term. TF-IDF typically performs better in machine learning models, even if Bag of Words vectors are simple to comprehend. The Snips NLU can be improved for higher probability in detecting intents by improved configuration of custom utterances in its .json file. These utterances

can also be borrowed from financial fraud texts (indicators) in prior legal cases that were investigated.

Few waypoints for fine-tuning the proposed financial fraud detection approach are 1) Use of training data in both quality and quantity. 2) Alternate labelling techniques to BERT if training data is unlabeled. 3) Customization of Snip NLU json for utterances. 4) Different set of Machine Learning algorithms can be used. 5) Variations in data preprocessing steps and 6) Suspect's risk ranking calculations.

```
## Create Dictionary {category:[keywords]}
print('Create Dictionary clusters for keywords Buy, Sell, Other ..')
dic_clusters = {}
dic_clusters["BUY"] = get_similar_words(['buy','purchase','acquire','invest','add'], top=30, nlp=nlp)
dic_clusters["Sell"] = get_similar_words(['sell','dispose','dump','unload','divest'], top=30, nlp=nlp)
dic_clusters["Other"] = get_similar_words(['amazon','android','app','apple','facebook', 'google','tech'], top=30, nlp=nlp)
```

Fig. 21. BERT logic for dictionary of clusters

Summary

In this chapter, an umbrella-approach is proposed that consists of multiple sub-approaches that together constitute a powerful tool for case investigators investigating financial frauds such as "Insider Trading" and "Pump and Dump". The combination of sub-approaches leverages automation, machine learning (supervised and unsupervised) algorithms, deep learning (transformers) techniques, risk profiling and suspect's sentiment analysis. Investigators can choose one sub-approach over another based on the results of each and the supporting indicators that they fetch. This research applies the sub-approach against synthetic case evidence dataset (ESI) that closely mimics real-world electronic case evidence such as Tweets, Facebook posts, emails, word documents, SMS texts and WhatsApp texts. These sources of data are notorious for deviating from traditional English

language and thus this research also highlights the need to address the linguistic challenges in case evidence before applying analytical techniques. The sub-approaches work on the intent of a suspect towards Internal trading and Pump and Dump (P&D) frauds. This research proposes pursuing the human intent during trading of stocks namely “buy” and “sell”. A suspect who is an employee of an organization (listed on the stock exchange) having privileged information from an event may trade stock (buy/sell) leveraging that privileged information for personal gain. While insider trading is not always a cause for concern, misusing company privileged information can be investigated and is punishable. The proposed approach can narrow down the electronic document (bates number) that exhibits an intent to “buy” or “sell”. This intent when coupled with the job title of the suspect, risk profile, access to the key events, etc. can assist case investigators in building winning legal arguments for the case. Likewise, a suspect exhibiting a pattern of intent through a series of “buy” followed by an intent to “sell” can be deemed as a P&D. Thus, the approach of pursuing intent in both the fraud scenarios can assist fraud investigators in pointing to the exact evidence (bates number) in the evidence pile thereby narrowing down the evidence and speeding up the investigative process.

CHAPTER VI

SEXUAL HARASSMENT DETECTION OF CASE SUSPECTS

Evidence for a real-life forensic investigation of sexual harassment was hard to find in public for academic research. This research customizes and builds random fictitious electronic evidence data (ESI) for this experiment [112]. This experiment was carried out using an Intel(R) Core (TM) i5-3470 CPU @ 3.20GHz 16 GB RAM PC and a 64-bit Windows 10 operating system. Software used was Python, PyCharm, SQL Server 2019, and Visual Studio 2019. This experiment showcases an umbrella-approach to identifying sexual harassment indicators from textual data in investigations. To avoid bias, the experiment results are published as-is, and no attempt was made to withhold wayward results or showcase only high-fidelity results.

Intents – Power, Persuasion, Abuse, Unwelcome and Humiliation

Human intents such as persuasion, display of power, abuse, unwelcome and humiliation were selected in this study as they are strong indicators of sexual harassment in conversations. Power, not lust, is considered the root cause of sexual harassment [125]. According to psychologists high-powered men accused of abusing women have different motivations, but often share some personality traits [126]. Sometimes, persuasion by predators can be more effective than force [127]. Although dating apps restrict persuasive attempts at contacting (dating) people, perpetrators can find means to approach the victim multiple times. As nouns there is a difference between harassment and abuse. Harassment is persistent attacks and criticism causing worry and distress while abuse is improper treatment or an unjust wrongful practice or custom. There is a thin line between abuse of the victim and sexual harassment, but any abuse with sexual overtones can be instrumental

in the investigation. Humiliation as an intent was chosen as sexual harassment usually leads to humiliation of victims threatening their physical and mental integrity [128], [129], [130]. Together these intents of a person's "mens rea" can help in the investigation as the intent is one of the two requirements that must be proven to secure a conviction (the other being the actual act, or "actus reus").

Experiment Design

The case evidence datasets used for this experiment were from prior research [112]. Key types of data were assembled from fictitious emails, Facebook posts, Tweets, WhatsApp/SMS messages, and random MS Word documents. Data was stored in SQL tables identified by their source/document identifier known as bates number/ID. Each email and MS Word documents were further broken into sentences and stored in a separate SQL table. Data needs to be processed for analytics as there can be occurrences of emojis, hyperlinks, stop words, etc. that can inhibit the analytical process [131]. All textual data was pre-processed using Natural Language Processing (NLP) techniques such as tokenization, stop words, stemming, and lemmatization. All suspect names, key event dates, textual data, and stock symbols used are solely for demonstration purposes and bear no resemblance in any shape or form in real life. For the machine learning and neural network models, this research undertook a three-pronged approach for labeled data and unlabeled data. A women's E-Commerce clothing reviews [132] dataset was used that contained reviews of women's dresses. The need for this dataset was felt appropriate as it largely commented on the looks of the person, dress colors, outfit sizes, and likes/dislikes. Such comments are largely found in sexual harassment scenarios and cyberbullying. Another dataset considered for the research was a labeled dataset ConvAbuse [133].

However, this dataset was largely unbalanced for the sexual harassment feature. Thus, another feature “type sexist” was combined with the “type sex harassment” feature as it closely aligns with sexual harassment. This labeled data was later compared against classification by BERT. Lastly, a sexual terminology lexicon dataset [134] was incorporated as many sexual harassment remarks can contain adult and vulgar words. Together, these datasets were used to identify intent from the evidence pile. Fig. 21 presents the overview of the experiment, and Fig. 22 presents the various sub-approaches (workflows) in this methodology to determine indicators of sexual harassment from textual evidential data.

Ancillary Data

For various automation steps, ancillary data such as emojis, emoticons, stop words, etc., were assembled from the Internet. Few data files were stored in SQL databases, while the rest were stored as flat files.

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for NLP was developed by Google [116]. BERT can be used in a wide variety of NLP tasks such as question answering and Natural Language Inference. A python script was written to classify women’s e-commerce clothing reviews [21] and ConvAbuse [22] datasets using sexual harassment intents such as persuade, power, abuse and humiliate. This approach helped in the unsupervised labeling of data. After text data preprocessing, creation of target clusters using Word2vec and gensim was performed, followed by word Embedding with transformers and BERT. The gensim package has a function that returns the most similar words for any given word.

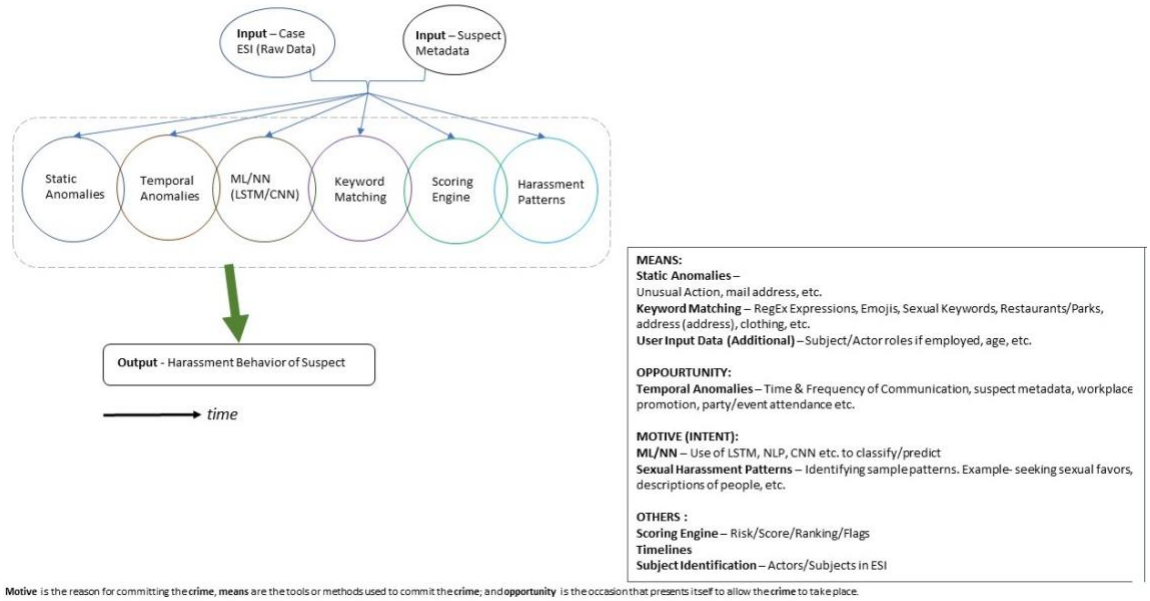

Fig. 21. Sexual Harassment detection – High level approach

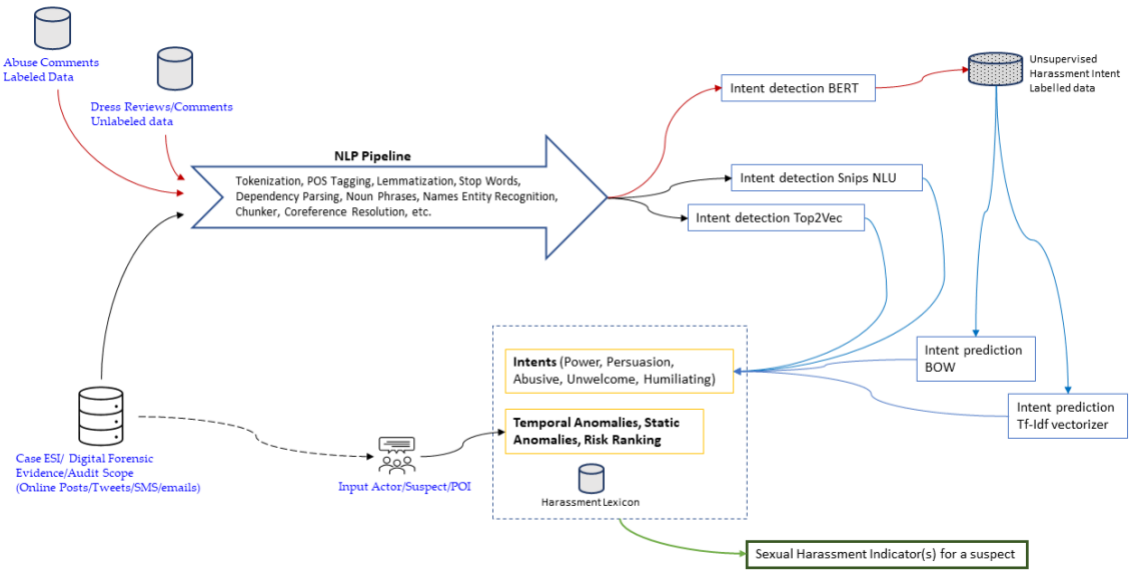

Fig. 22. Sexual Harassment detection process involving multiple approaches

Lastly, assignment of observations to clusters were done using cosine similarity and the model's performance was evaluated. Classification results were stored in SQL tables as labeled data using the BERT approach.

Term Frequency – Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical approach that determines how important a word is by weighing its frequency of occurrence within the document. After data preprocessing (word cleaning, stop words removal, hyperlinks, stemming, lemmatization), the data from earlier used BERT technique was split into training & testing subsets. The training data was fitted with a Naive Bayes classifier, and predictions were derived using the test dataset. Model's accuracy, precision, recall, confusion matrix, and ROC were obtained. This trained model was then applied against textual evidence (identified by bates number) from the investigation caseload. Prediction results of persuasion/power/abuse/humiliate intents were stored in a SQL table.

Snips

Snips NLU [135] is an open-source Natural Language Understanding (NLU) python library that allows for parsing sentences written in natural language, and then extracting structured information [119]. The NLU engine first detects the intention of the user from the text using custom utterances defined in a json format. This json was then fitted into the SnipsNLU Engine with persuade/power/abuse/humiliate intents. An excellent alternative to Snips NLU was Rasa NLU [120]. However, Snips NLU was proven better than Rasa NLU [136], [122] and was thus used in this experiment. Results were stored in SQL tables.

Suspect Metadata & Risk Profile

Suspect metadata can provide valuable information. Harassers can carefully build up an image so that people would find it hard to believe they would do anyone any harm. There are many types of sexual harassers like power-players, serial harassers, gropers,

opportunists, bullies, pest, confidante, situational harassers, stalking, intellectual seducer, great gallant, and mother/father figure (the counselor-helper) [137]. In this experiment, the investigators store metadata information of suspects collected during the investigation. This metadata can then be used to calculate the risk profile of the suspect using a weighted approach.

Presentation

A Windows Forms (client/server) software module was created using C#.NET as a custom prototype tool for use by the investigators. Fig. 23 shows the custom tool screen developed for case investigators. The custom tool inputs the suspects from a case investigation, executes Python and C# scripts and outputs the bates number that contains indicators of sexual harassment for the selected suspect.

Analysis

This research combines supervised learning and unsupervised learning to identify indicators of sexual harassment from synthetic digital forensic case evidence (ESI). The algorithms - BERT and Snips - used in this research were chosen as they work well for NLP, NLU and can be further improved with user feedback (fine-tuning) via the custom software developed for this research experiment. This approach is suitable for any investigation team that has no prior labeled data on sexual harassment intents. They can start with unlabeled data and, over the period of a few investigations, build a quality labeled dataset.

Quality of Digital Evidence

The quality of forensic data is an essential component of machine learning algorithms. Digital forensic data (evidence) in a legal case should be carefully worked on

to preserve its integrity. Any violation of integrity can render it inadmissible in court. Unfortunately, data from the Internet world can contain a lot of slang, jargon, abbreviations, emojis, gifs, emoticons, smileys, hyperlinks, and typos. Digital forensic evidence can contain data from the Internet and thus need some level of cleaning before use in machine learning and deep learning algorithms. Data cleaning is the process of preparing raw text for NLP (Natural Language Processing) so that machines can understand human language. Applying indiscriminate data cleaning steps to obtain a better data quality for analysis can be detrimental to the interpretation of the original textual evidence. When using the BERT approach, similar words for word vectors should be carefully planned. Likewise, the utterances in snips should not introduce bias. Duplicate/irrelevant data may be ignored, but missing data should not be added back in. Any corrupted data or outliers should be skipped. To summarize, depending on the evidence text being cleaned, the output of analytical algorithms can vary but, in doing so, may alter the evidence during analysis, making it and the analysis results inadmissible in a court!

Supervised/Unsupervised Learning

The datasets used in the experiment were randomly picked and assembled to mimic typical digital forensic case evidence and investigation. Twitter data, WhatsApp data, SMS data, emails, random custom MSWord documents, and Facebook data constitute the case evidence. Thus, the accuracy of models and results of the experiments were solely for demonstration of the approach. Using the BERT approach, the ConvAbuse dataset was re-labeled for sexual harassment indicators. BERT model achieved a 54% accuracy when predicting sexual harassment intents (power, abuse, persuade, unwelcome, humiliate). This

way, a previously labeled dataset can be re-labeled using BERT with custom dictionary cluster keywords. The cluster keywords chosen for this research are shown below and were top occurrences of the previously labeled sexual harassment data of this dataset.

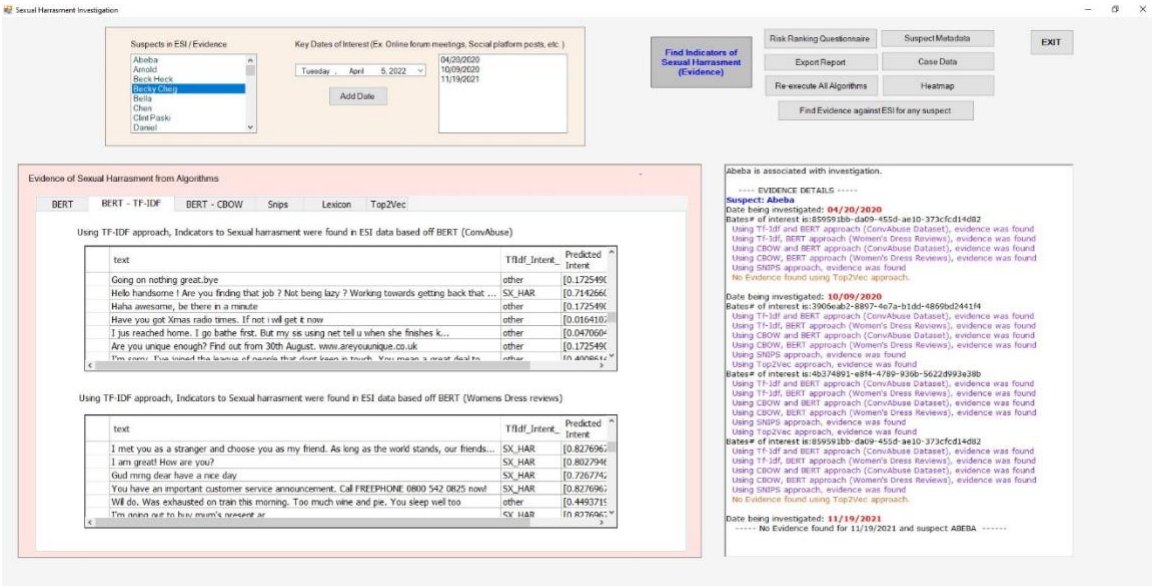

Fig. 23. Custom application screen identifying sexual harassment indicators of a suspect found from synthetic digital evidence [Note: Names shown are purely for academic study and have no bearing on an event or person or investigation.]

The investigation team can alter these keywords as needed to label any historical case data. Fig. 24 shows the ROC and precision of using BERT to re-label the ConvAbuse dataset for sexual harassment indicators. Any change of parameters such as the number of clusters and the list of similar keywords can greatly impact the results and classification accuracy. Parameters to create a cluster dictionary in the BERT code logic is shown below.

```
dic clusters["SX HAR"] = get similar words(['abuse', 'power',
'persuade', 'unwelcome', 'humiliate', 'strength', 'exploit',
```

```
'cajole', 'exploit', 'dick', 'sex', 'horny', 'love'], top=30,
nlp=nlp)

dic clusters["other"] = get similar words(['please', 'flying',
'city', 'sure', 'offset', 'flight', 'tech', 'buy', 'sell',
'seasons', 'gas', 'greenhouse', 'emission', 'project'], top=30,
nlp=nlp)
```

This re-labeled data from the BERT approach was then used by TF-IDF and BOW to predict sexual harassment indicators against each tweet or each Facebook post in the evidence pile. The emails and word documents in the case evidence pile were parsed by the BERT model for each sentence.

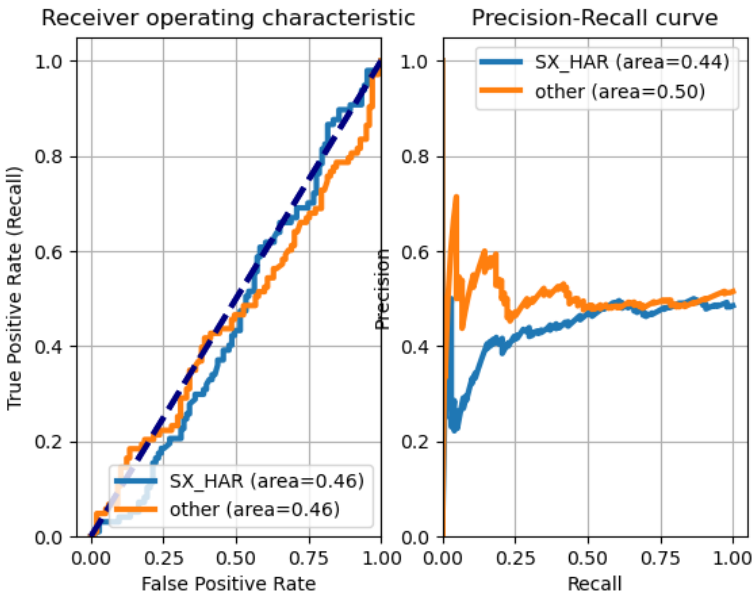

Fig. 24. BERT: ROC & Precision recall

TF-IDF achieved an 85% accuracy while BOW achieved a 86% accuracy in prediction. This data was displayed by the custom tool developed. The investigators can discount any inconsistencies by the tool feedback process. Fig. 25 shows the ROC and precision of using BOW and Fig. 26 shows the ROC and precision of using TF-IDF using the BERT labeled ConvAbuse data for sexual harassment indicators. The accuracy in categorization by BERT directly impacts TF-IDF and BOW classification accuracy.

For the women's clothing reviews dataset that was unlabeled, BERT approach was applied to label the data for intents (power, abuse, persuade, unwelcome, humiliate). BERT model achieved a 53% accuracy when predicting sexual harassment intent. This way, a previously unlabeled dataset can be labeled using BERT with custom dictionary cluster keywords. The BERT logic cluster keywords chosen for this research are shown below and were top occurrences of dataset alluding to sexual harassment or otherwise. The investigation team can alter these keywords as needed to label any historical case data.

This labeled data using BERT approach was then used by TF-IDF and BOW to predict sexual harassment indicators against each tweet or each Facebook post in the evidence pile. The emails and word documents in the case evidence pile were parsed by the BERT model for each sentence. TF-IDF achieved an 87% accuracy while BOW achieved an 85% accuracy in prediction. This data was displayed by the custom tool developed. The investigators can discount any inconsistencies by the tool feedback process. The accuracy in categorization/labelling by BERT directly impacts TF-IDF and BOW classification accuracy.

Using the Snips approach, json files with utterances were created for each of the intents (power, abuse, persuade, unwelcome, humiliate). A python script was employed to

apply snips NLU engine against the forensic case data. Sexual harassment indicators on text evidence data for each intent were observed with accuracy. The intent “persuade” was identified by snips with a 48% accuracy, “abuse” with a 55% accuracy, “humiliate” with a

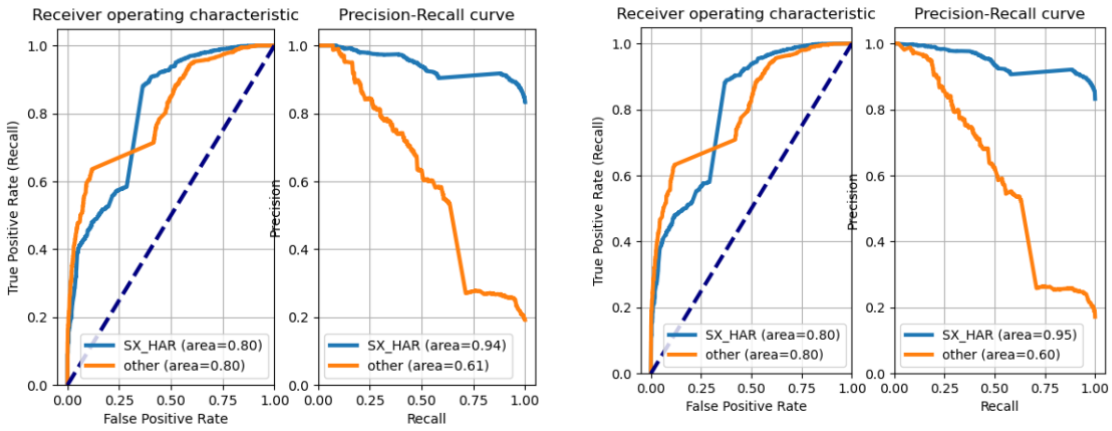

Fig. 25. BOW - ROC and Precision Recall of Women’s Clothing reviews (L) and ConvAbuse (R) when labeled using BERT

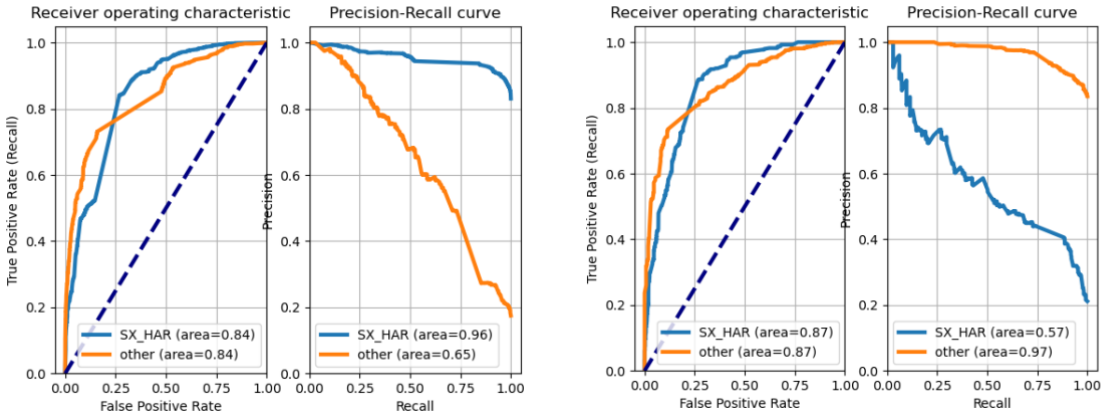

Fig. 26. TF-IDF - ROC and Precision Recall of Women’s Clothing reviews (L) and ConvAbuse (R) when labeled using BERT

29% accuracy, “power” with a 42% accuracy and “unwelcome” with a 50% accuracy. Fig. 27 shows the sample utterances used. The accuracy of the snips NLU engine is highly dependent on the utterances in the YAML input file. Any utterances that can be verified as sexual harassment can be used.

```
type: intent
name: power
utterances:
  - why did you not complete the work?
  - meet me for our project updates at dinner
  - I need to you to add this to my portfolio
  - show me gains for our stock this evening
  - do it or else
  - This is the last time you will mention the incident
  - You can stay back in the office when we go out
  - I have access to your door keys
  - My boss ordered me to organize tons of documents that I would never be able to finish within my working hours.
  - I was told not to leave until I finished.
  - I was forced to take up the playground space in the morning
  - I was not given any work that I could use my experience for
  - I was ordered to do simple tasks endlessly.
  - My boss didn't give me any work at all.
  - I was sitting in front of the phone every day as a punishment for nothing
  - I was removed from the team when I contradicted my supervisor.
  - My coworkers began to ignore me in a group
  - Emails and necessary information that should be shared with everyone were never sent to me
  - I was told to wear a skirt at work
  - I was scolded every day in front of other colleagues
  - He will always walk by my front yard
  - She would send me out on silly errands
  - He liked to come close to me at work
  - He feels like a boss
  - My boss checks my phone whithout my permission
  - He told me to hush up about his illegal acts
```

Fig. 27. Snips YAML logic containing sample sexual harassment utterances

The forensic investigation case data (evidence) was categorized by Top2Vec [117] algorithm for topics similar to the intents (power, abuse, persuade, unwelcome, humiliate). Few matches to intent “power” were observed but, upon manual review, were found to be incorrectly flagged. This can be attributed to the data cleaning steps employed that adversely impact topic categorization. The results from Top2Vec were displayed in the custom tool developed. The case investigators can discount such inconsistencies by using the tool feedback process.

A lexicon dataset [23] was used to flag exact keyword matches and similar words against the text evidence data. The results were displayed by the custom tool developed. The case investigators can tweak such lexicons for pattern matches and similar words. This process can further assist investigators in drawing conclusions in addition to the other analytical approaches mentioned.

To summarize our experiment metrics, the BERT model achieved a 53% accuracy in predicting sexual harassment intent. TF-IDF model prediction accuracy was 85% and

the BOW model prediction accuracy was 86%. The probability of Snips NLU in determining intents: “persuade” was 48%, “abuse” was 55%, “humiliate” was 29%, “power” was 42% accuracy and “unwelcome” was 50%.

Based on the idea of Machine Learning Transformers, Google's BERT considers entire sentences rather than just individual words for various language tasks. BERT is regarded as the state-of-the-art language model for NLP and was specifically trained on Wikipedia (about 2.5 billion words) and Google’s Books Corpus (about 800 million words). BERT leverages Masked Language Model (MLM) by enabling/enforcing bi-directional learning from the text by masking (hiding) a word in a sentence and using the words on either side of the masked word to predict the masked word [124]. In our use of BERT for unsupervised labelling of sentences for intents, the BERT code can be further fine-tuned for better accuracy when defining a dictionary of clusters. Fig. 28 displays the code used for defining the clusters in our BERT logic. The price of this fine-tuning can be higher CPU usage and longer application execution times. Bag of Words (BOW) and TF-IDF (Term Frequency Inverse Document Frequency) are the two models that can be used after labeling to predict financial fraud indicators. BOW generates a collection of vectors that count the number of times each word appears in the document or sentence. The bag-of-words model is frequently employed in document classification techniques in which the (frequency of) occurrence of each word is used as a feature for instructing a classifier. The TF-IDF model includes information on both significant and uncommon terms, and it assigns higher values to uncommon words in the corpus of documents. The TF-IDF score rises when a word occurs more frequently to account for the fact that some words are used more frequently than others overall. TF-IDF typically performs better in machine learning

models, even if Bag of Words (BOW) vectors are simple to comprehend. By modifying the configuration of custom utterances in its .json file, the Snips NLU can also be modified for a higher probability of identifying intentions. These utterances can also be borrowed from sexual harassment texts (indicators) in prior legal cases that were investigated.

A few waypoints for fine-tuning the proposed sexual harassment detection approach are; 1) Use of good quality and quantity training data. 2) Use of alternate labelling techniques to BERT if training data is not already labelled. 3) Customization of Snip NLU .json for utterances 4) Alternative Machine Learning algorithms used. 5) Variations in data preprocessing steps and 6) Suspect's risk ranking calculations.

```
## Create Dictionary {category:[keywords]}
print('Create Dictionary clusters for keywords ..')
dic_clusters = {}
dic_clusters["SX_HAR"] = get_similar_words(['abuse','power','persuade','unwelcome','humiliate','strength','exploit','cajole','exploit',
                                            'sex','horny','love'], top=30, nlp=nlp)
dic_clusters["other"] = get_similar_words(['please','flying','city','sure','offset', 'flight','tech','buy','sell','seasons','gas',
                                           'greenhouse','emission','project'], top=30, nlp=nlp)
```

Fig. 288. BERT logic dictionary of clusters

Timeline is a key factor in all investigations and similarly a timeline of harassment indicators can help with legal arguments. The custom software helps with plotting a timeline of sexual harassment indicators of the suspect. The risk of a suspect exhibiting sexual harassment behavior was calculated based on indicators found and his/her risk profile. The risk questionnaire consisted of attributes collected such as gender, age, job title (workplace scenario), conversation in online setting (yes/no), prior offences/strikes of harassment behavior, etc. These attributes/metadata can then be ranked with weights to arrive a risk level. Certain items on the risk questionnaire can carry additional weight than others, for example, there are higher chances of harassment behavior exhibited online than in-person conversation as perpetrators can get away with online anonymity. The

investigators can use this risk ranking logic along with indicators found using the analytical approaches to justify the suspect's exhibited sexual harassment behavior.

Summary

Identifying indicators of sexual harassment from written textual evidence of an investigation can be challenging. Evidence in a case typically constitutes social media data, emails, and text messages. In a workplace setting, sexual harassment can be found in Microsoft Office documents such as memos or termination letters. These sources may sometimes offer poor quality of language data as Internet users may insert slang, typos, emojis, etc. Also, the unavailability of quality labeled sexual harassment data for supervised learning can be an impediment to investigators in leveraging analytics and NLP. This research proposes an approach that consists of multiple sub-approaches that together constitute a powerful tool to identify sexual harassment indicators from textual digital forensic evidence. The proposed solution addresses the lack of labeled data specifically for sexual harassment indicators.

CHAPTER VII

ANALYTICS IN DIGITAL FORENSICS AND EDISCOVERY SOFTWARE

With the recent rave in analytics, such as Artificial Intelligence (AI), Machine Learning (ML), Neural Networks (NN), and Deep Learning, leveraging these techniques into custom digital forensics or eDiscovery software to analyze case evidence has been highly beneficial. Analytics, together with automation, has helped reduce investigation time and thereby costs. With typical case evidence data volumes every increasing due to affordable Cloud storage and cheap smartphones, mining of evidentiary data for clues and indicators to support legal arguments has become a mammoth task. Incorporating analytical approaches into digital forensic or eDiscovery software to speed-up case investigation with quality results is thus a focal area in software product development and academic research. This chapter also touched upon the challenges and opportunities faced when leveraging analytics in custom forensic software development and its use.

Analytics – Null Data

Evidentiary data can contain Null values in certain cells or for the entire row. A NULL value is a placeholder to denote values that are missing. Comparisons and arithmetic operations with a NULL produce NULL results and are thus meaningless to analytical techniques. NULL values in digital forensic data generally fall into one of two categories: values that are missing at random due to limitations of the forensic software collection mechanism and those values that are not missing at random due to design flaws of the electronic device. For example, digital forensic software extracting web browsing history may report a NULL value on a hyperlink as the browser allows for NULL hyperlinks under bookmarks, or a caller name is NULL in phone contacts of a smartphone as the device

allows blank names on the contacts. If the field is allowed to be blank/NULL on the device by design, data extraction by forensic software can report it as blank/NULL. While there is a difference between blank data and null values, there a possibility that null values may exist in data collected from forensic devices. The challenge lies in how to now process null or blank data during analytics as we cannot ignore rows with null as it can result in the filtering of evidentiary data tantamount to accomplishing our goals by introducing bias. This is a problem in historical data used for analytical learning as well as during the analysis of evidentiary data at hand. As a digital forensic analyst, one should always check evidentiary data with a histogram for NULL values, blank rows or cells, null reported in a string format, and occurrences of "N/A". Ignoring all the rows containing a NULL value might not be a wise decision. Instead, prior to applying any analytical processes to the evidentiary data, it is advisable to document in detail the cells or rows (with NULL or blanks) that are 1. Missing completely at random (MCAR), 2. Missing at random (MAR) and 3. Missing not at random (MNAR). The implications of NULL values missing completely at random MCAR) can be catastrophic for the validity of the analysis techniques, investigation, and case arguments. Further attempts at forensic data extraction from the electronic device may address MAR and MNAR. If a decision is made to ignore rows or cells with NULLs, adequate documentation must be made to explain what the analysis result would be if such data was used and how in turn it would impact the result.

Repeatability, Randomness, and Sampling

According to the National Institute of Standards and Technology (NIST) NISTIR 8006 [138] and Digital Forensic Research Workshop (DFRWS) [139], forensic test results must be repeatable and reproducible to be considered admissible in a legal setting. Digital

forensics results are repeatable when the same results can be obtained repeatedly when using the same methods as in the same testing environment. In analytics, data preparation requires splitting a dataset (evidence data) into training and testing for supervised learning. We should avoid randomization before splitting train and test datasets as each run of the experiment will then yield different results due to the randomness involved in data selection. Seeding ensures that the Random Number Generators (RNG) output the same values in the same order each time we run it, recreating the dataset [140]. Hashing is a common way to split or sample data; however, the inputs to our hash function should not change each time we run the data generation program. Lastly, use of the current time or a random number as inputs to the hash should be avoided if we want to recreate our hashes on demand or replicate our experiment.

Class imbalance may affect our evidence datasets with more than two classes that may have multiple minority classes or multiple majority classes. Data sampling provides a collection of techniques that transforms a training dataset to balance the class distribution [141]. Oversampling or undersampling should be avoided in an imbalanced class distribution. Oversampling methods duplicate examples in the minority class or synthesizes new examples from the examples in the minority class [140]. Duplication of evidence data for the sake of arriving at results in the investigation should be avoided as it interferes with the state and integrity of the case evidence.

Reporting, Logs, and Audits

Digital forensic software leveraging analytics must have plenty of visualization features like heat-maps, graphs, and charts. To the jury or at the court, statistical graphs such as for ROC, AUC, precision-recall, or accuracy may be of limited use, but rather the

people/jury in the courtroom would like to see graphs and charts that they can easily infer from. Traditional reports are also encouraged, along with data exports and drill-down reports. Logs to support an audit trail are a must as part of the repeatability requirement of digital forensics. If needed, other investigators must be able to follow the logs and trigger actions on the software/software to reproduce the same results. Lastly, digital forensic software supporting analytics must allow for audits and offer an audit role type of user access with restricted access privileges.

Date & Time Format

Dates and time feature data are critical to a forensic investigation, and their formats can be detrimental to the success of the case. Care should be taken to first convert/encode all date and time data into a specific time zone and then apply proper conversion techniques. For example, pandas views date time data as strings. To convert these strings into datetimes (datetime64), we should use the pandas function to datetime along with the format parameter and convert errors into not a datetime (NaT).

Data – Warehouse Or Database

A database is an organized collection of information stored in a way that makes logical sense facilitating easier searches, retrieval, manipulation, and analysis of data. Databases can be either SQL or NoSQL based. SQL based databases can scale vertically, while NoSQL can scale horizontally. SQL (relational) databases are less flexible and more rigid in terms of the data hierarchy but support queries that are easy to use and can be tuned for performance. A data-warehouse is a system that aggregates and stores information from a variety of disparate sources. Data-warehouses are designed from the ground up mainly for reporting and analysis purposes. True and verified copies of case evidence data can be

imported into databases or a data-warehouse. The question arises as to which is the best data storage option for analytical experiments. As analytical experiments grow using the database, managing schema objects can get complicated, requiring additional database administrator resources to manage the database. Similarly, a data-warehouse may seem to be a design overkill but can scale better when multiple analytical experiments are being conducted. However, data-warehouses do not support multiple concurrent connections as databases do. Storing case evidential data as a flat-file for analytical experiments is not advisable as flat-files do not support complex searches and read/write transactions as robustly as databases or data-warehouse.

Privacy PHI/PII in Evidence Data

Case evidence can contain Protected Health Information (PHI)/Personally Identifiable Information (PII)/Confidential Business Information (CBI) data causing data privacy and access concerns in handling of such data during analytical experiments. Also, to arrive at a good quality of training data for supervised learning, sometimes historical case investigation data may be a good source to start with. However, using such historical data for analytical research can also raise legal concerns of privacy and ethics. While such historical legal cases may be closed and now archived, ownership of such data, and reuse of it to build a training dataset may itself need legal, privacy and client approvals. For example, to build a training dataset for Facebook posts containing financial fraud evidence, the analytical team may want to tap into forensic evidence data from historical cases. In such instances, the ownership of the historical data and related privacy concerns of its use will need to be clarified.

Encryption in Evidence Data

Sensitive evidence data that was stored in an encrypted way will need to be decrypted for use in analytical experiments. This leaves the data in an unsecured state, and care must be taken to re-encrypt it at the earliest. Likewise, results of analytical experiments utilizing this decrypted data may in-turn contain data that now needs to be secured. Allowing the analytical team to access the keys to decrypt and re-encrypt sensitive data can be a security risk.

Verification and Validation

Analytical methods and models used in digital forensics to analyze/mine case evidence can be called into question and opposed in courts. The challenge arises in the experiment/ model/method verification and validation process. To better understand this challenge, we need to understand the types of data used in an analytical experiment.

1. Training data - This type of data helps build the machine learning algorithm within the analytical experiment. Data is input to the machine learning algorithm resulting in the expected output. The model repeatedly evaluates this data to learn more about the data's behavior and then adjusts itself to serve its intended purpose [142].

2. Validation data - During model training, new data can be infused into the model as part of validation. This new data is known as validation data or holdout set and is often 10% of the total data which was not used by the model as yet. Validation of data can be tricky as it requires significant understanding of the data in order to select the correct approach such as k-fold cross validation or time-based splits. Validation data provides the first test against unseen data, allowing the forensic team to evaluate how well the model can make predictions based on the new validation data. The use of such validation data is

uncommon but advised in a forensic analytical experiment as it can provide helpful information to optimize hyperparameters, which influences how the model assesses data [142].

3. Test data - After the model is built, trained, and validated, testing data once again validates that the analytical model can make accurate predictions. The testing data should be left unlabeled if the training and validation data included labels to evaluate the model's performance metrics. Test data is a last, real-world verification of an unknown dataset to ensure that the machine learning algorithm was properly trained [142].

Thus, utilizing validation data in the analytical experiment can provide an initial check that the model can return useful predictions in a real-world setting, which training data cannot do. Validation data can be part of the training data but is advisable to be an entirely different dataset than the training dataset [142]. The use of validation data can also reassure the jury or the court that the model's algorithm works as intended in predicting results as part of the analytical experiment.

Metrics and Graphs

Analytical experiment results are best represented in graphical formats along with key metrics such as model accuracy and loss. To be well understood and accepted in a court or by a jury, visualization of analytical experiment's decision-making process results such as evaluation metrics, learning curves, scatter plots, performance charts (like ROC, Lift Curve, Precision-Recall charts, confusion matrix, etc.) is critical. Further use of visualization techniques to summarize the investigation focus and analytical experiment results is advisable. For example, a bar chart on instances of sexual harassment indicators by the suspect over a period of time can be added on top of the analytical experiment's

model prediction accuracy and precision-recall chart. Care must be taken to not over-burden the jury or court with statistical graphs, model architecture, and detailed metrics unless called for.

Domain Ontology Limitations

In the case of large volumes of data, automation coupled with data mining and AI can greatly speed up the forensics process and thereby allow for a quicker investigation. However, decisions made by and with the assistance of AI based forensic software need to be justifiable and explainable to a jury. Often, analytical experiments, AI algorithms, and accompanying automation tend to be too scientific for lay people and thus EXplainable artificial intelligence (XAI) will need to be employed wherein lay explanations for outputs are provided when leveraging analytics [143]. As AI technology and capabilities advance over time, it may become more difficult, or even impossible for AI systems to be explainable to a jury or in a courtroom. Thus, care must be taken during courtroom evidence presentation to limit results to simple graphs/ charts, metrics, graphical execution plans, drill-down reports, etc. from forensic software leveraging AI and from analytical experiments conducted on case evidence.

Multiple Analytical Approaches

The design of digital forensic software supporting analytics should involve multiple approaches and allow the user (investigator) to choose the most appropriate one. For example, if the custom forensic software addresses multi-class classification, multiple algorithms such as k-Nearest Neighbors, Decision Trees, Random Forest, Gradient Boosting and Naive Bayes may be offered by the software thereby allowing the user (investigator) to choose the most appropriate one based on classification results. This way,

the software does not limit itself to one approach/algorithm but rather offers a variety. Limiting to one algorithm may prove detrimental as a specific algorithm may not work best across multiple datasets (different case evidence data).

Security - Access Control, Evidence Destruction

While the case investigator may enjoy a certain degree of his/her access to the current case evidence on hand, their access to certain historical case data or labeled data will need to be considered. Case evidence may have PHI, PII, or CBI data making privacy and security key aspects of any analytical experiment. Disseminating results post-analytical experiments may need such results to be circulated and stored with colleagues or shared with clients. This would call for triggering necessary data privacy and security access controls to both experiment results and other automation logs. All case evidence must have an end-of-life timeline defined. Analytical research experiments using historical or ongoing case evidence must factor these timelines as the results of these experiments themselves may contain copies of the original evidence. Uncontrolled sharing of these results can also lead to complications to evidence destruction.

Software Development

The custom software/software developed for this experiment was developed using an Intel(R) Core (TM) i5-3470 CPU @ 3.20GHz 16 GB RAM PC and a 64-bit Windows 10 operating system. Software and programming language used was Python, PyCharm, SQL Server 2019, C#, and Visual Studio 2019. The custom software “Digital Forensic Case Evidence Analytics” (DFCAE) supports multiple modules such as suspect’s sentiment analysis, financial fraud indicators of suspects, and sexual harassment indicators - all leveraging automation, data mining, and analytics. The software user interface was

written using C#, calls necessary Python files and stores all data on a back-end SQL Server database. For software to be deployed and used by digital forensic and eDiscovery professionals, the authors decided to use WinForms and ultimately develop a client/server based Windows executable file with supporting DLL files. Each case evidence has its own database, and a common database serves as a master repository for labeled/unlabeled training data for analytics. Fig. 28 shows the growing complexity of traditional database schema objects for storing forensic evidence when used for analytic experiments.

Text from case evidence is mined using analytics and automation for results (indicators). For each upload, a new database in the SQL server instance is created and a few schema objects are automatically defined as part of SQL scripts. The software can handle evidence data from sources such as Facebook posts, Twitter data, SMS/WhatsApp messages, emails, and MS Word documents. Case investigators can switch between the three modules (Sentiments of suspects, Financial Fraud Detection of suspects and Sexual Harassment Detection of suspects) against the same case evidence. This way, the investigators have a choice to pursue different investigations against suspects of the case from the evidence collected. The investigators can store case metadata, upload evidence, review evidence statistics, trigger sentiment analysis, search for indicators of financial fraud and find indicators for sexual harassment. Fig. 29 and Fig. 30 show the user interface screen for the investigation case metadata and ESI (case evidence) metadata. While each of these key features has been discussed in detail in previously published articles of this project [112], we will briefly touch upon them.

The sentiment analysis of suspects found within the case evidence is carried out using multiple approaches and algorithms. Fig. 31 shows the user interface module to

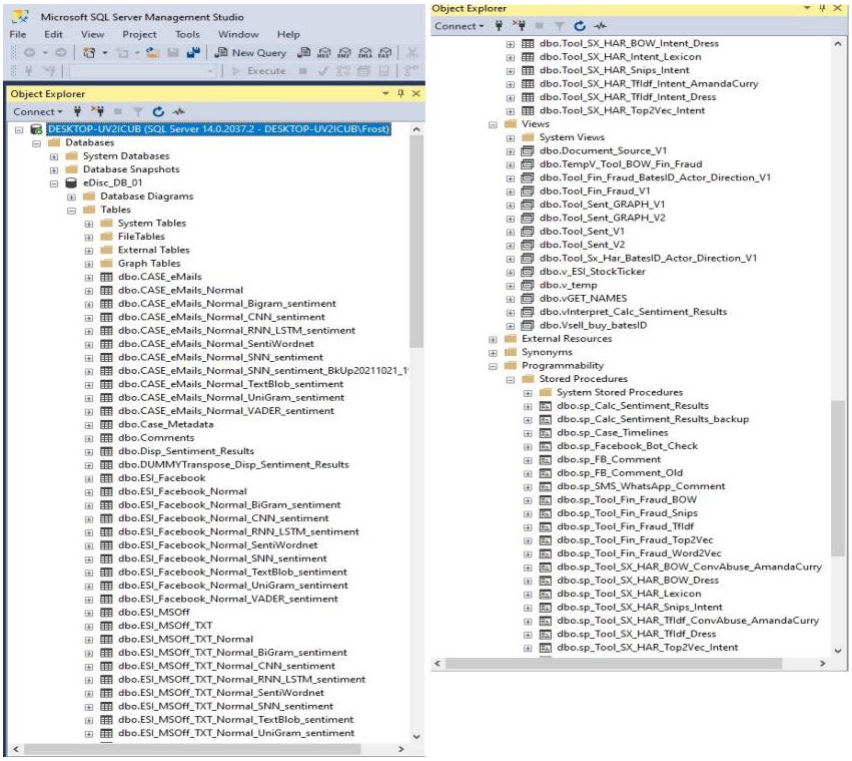

Fig. 299. Database Schema view of custom forensic analysis software showing complexity of database schema and design when using a traditional database

detect sentiments of case suspects from case evidence. Thus, the investigators can trigger the module for multiple analytical approaches. The results are then displayed on the user screen. Currently the sentiments are either positive or negative but can be scaled depending on the training data that is uploaded via the software. Investigators can access various reporting functions like charts, heat-maps that point to evidence source and the sentiments of the suspect. Results can be exported to a flat file. Prior work on the detection of sentiments of case suspects using this custom software is available on GitHub [112].

A financial fraud detection module detects fraudulent behavior in pump and dump schemes and insider trading using multiple analytical approaches. Fig. 32 shows the user interface module for detecting financial fraud indicators from case evidence. The

investigators can choose stocks to target and the suspect of interest. The module predicts from evidence the sources that have strong indicators of such financial fraud. The module correlates to historical stock data from Yahoo Finance. Investigators can access various reporting functions like charts, heat-maps that point to evidence-source and the fraudulent behavior of the suspect. Results can be exported to a flat file. Prior work on the detection of financial fraud of case suspects using this custom software is available on GitHub [112].

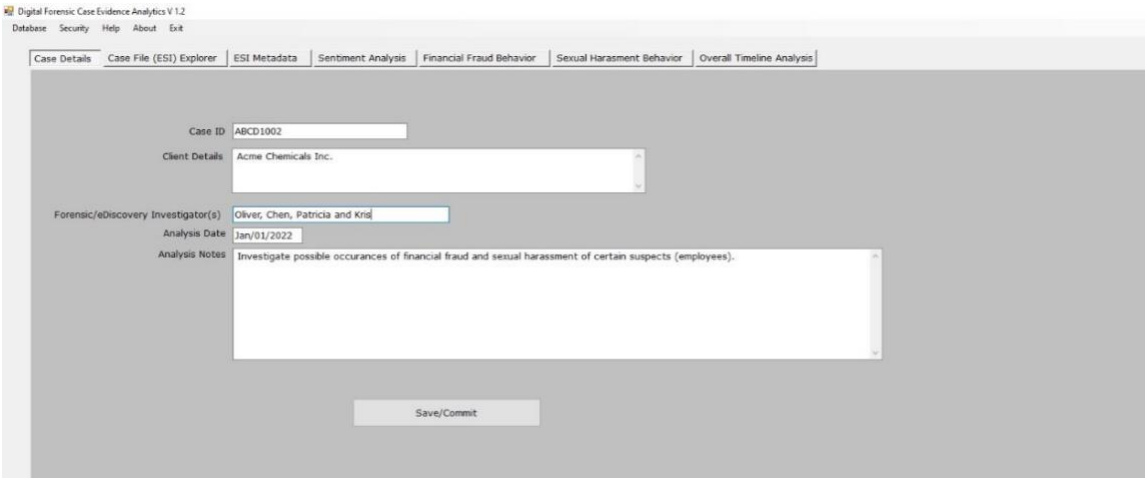

Fig. 30. Screen capture of case metadata on the custom software

The sexual harassment detection module detects possible sexual harassment of a suspect using multiple analytical approaches. Fig. 33 shows the user interface module for detection of sexual harassment from case evidence. The investigators can choose a suspect and trigger the module to predict indicators of possible sexual harassment. Investigators can access various reporting functionality like charts, heat-maps that point to evidence-source and the harassment behavior of the suspect. Results can be exported to a flat file. Prior work on the detection of sexual harassment indicators of case suspects using this custom software is available on GitHub [112].

This software allows for logging all user activity thus allowing for future audits. These activity logs are stored in the database and can be exported when needed for software usage audits or for training. Each analytical program triggered also contributes its run-time debugging information to a run-log flat-file, which can be accessed from this software. The software allows for further insight into case evidence by displaying a HTML based time-series graph using Google's API for charts. Fig. 34 showcases the communication timelines (from case evidence) of case suspects in our custom forensic analysis software. A help module was also created for the software, along with a security module was created for role-based access for users of this software. The source code, along with this software project files and repository, can be accessed online on GitHub [112].

Summary

Designing and developing digital forensic software that leverages analytical techniques requires careful design and back-end planning. The design of such software should factor in logging, security, and privacy requirements. Investigators would often need multiple analytical approaches from the custom forensic software to choose the best model. In this paper, the authors discuss a custom forensic software developed for multiple use cases. The authors also discuss best practices in developing such custom forensic software that supports analytics. As part of future work, the authors plan on adding additional modules such as stenography detection and signature detection while expanding its support for ingesting web-browser data and Portable Document Format (.pdf) files from the case evidence.

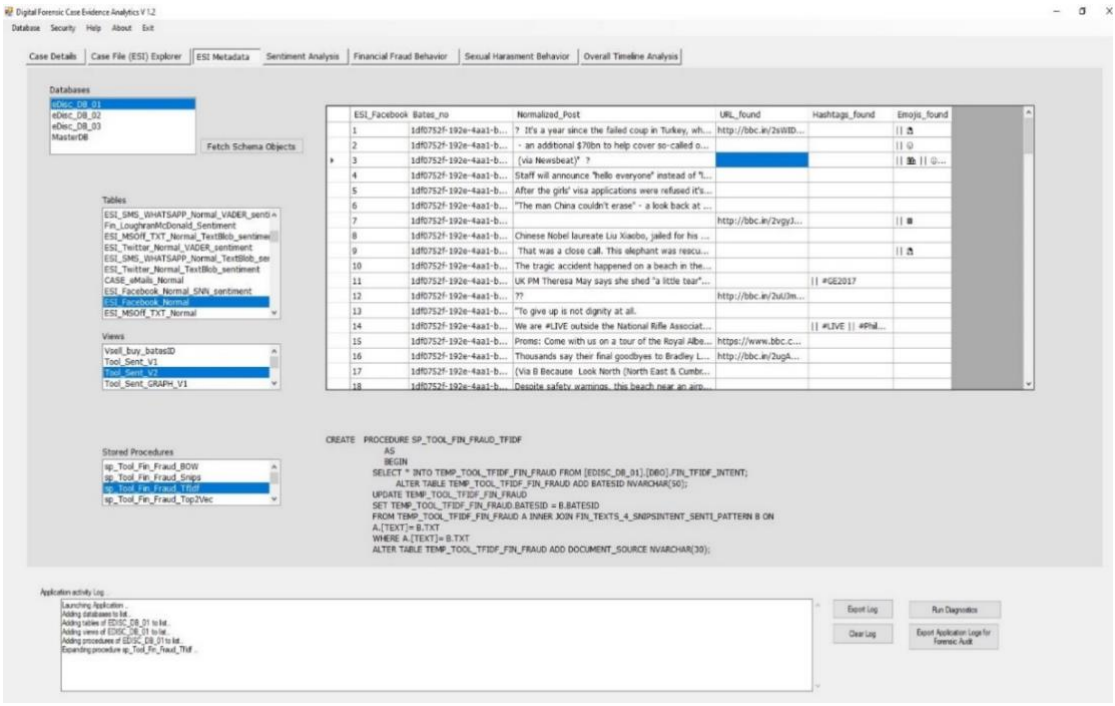

Fig. 301. Database schema view of the custom software for each case evidence (ESI)

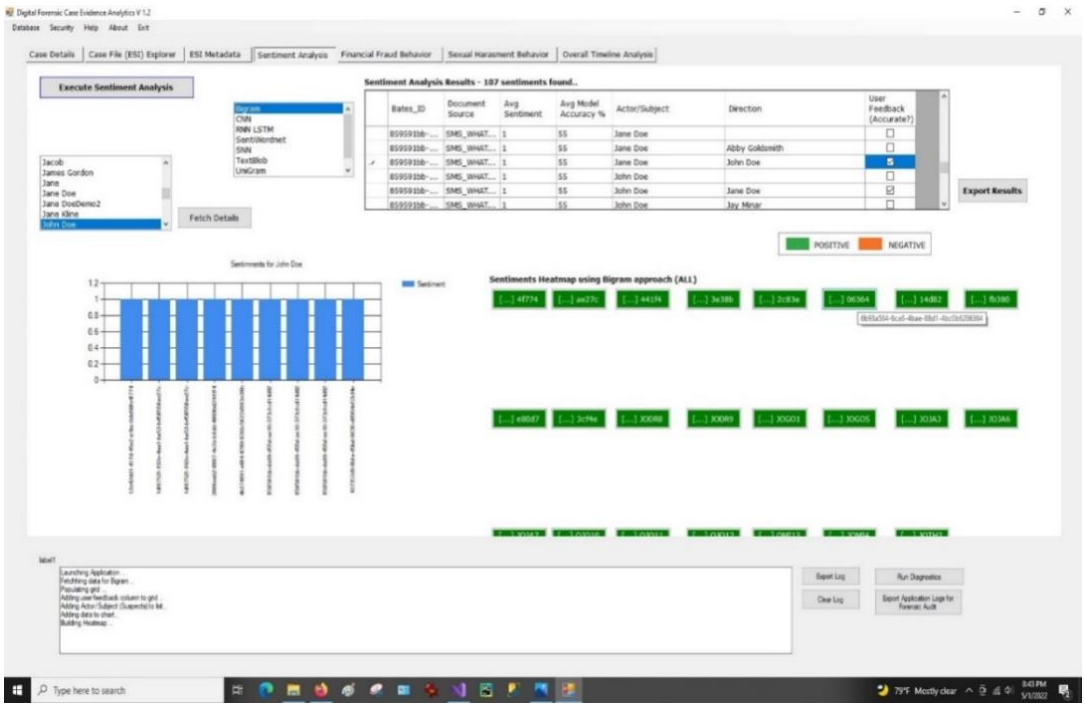

Fig. 312. Sentiments of case suspects using the custom forensic analysis software

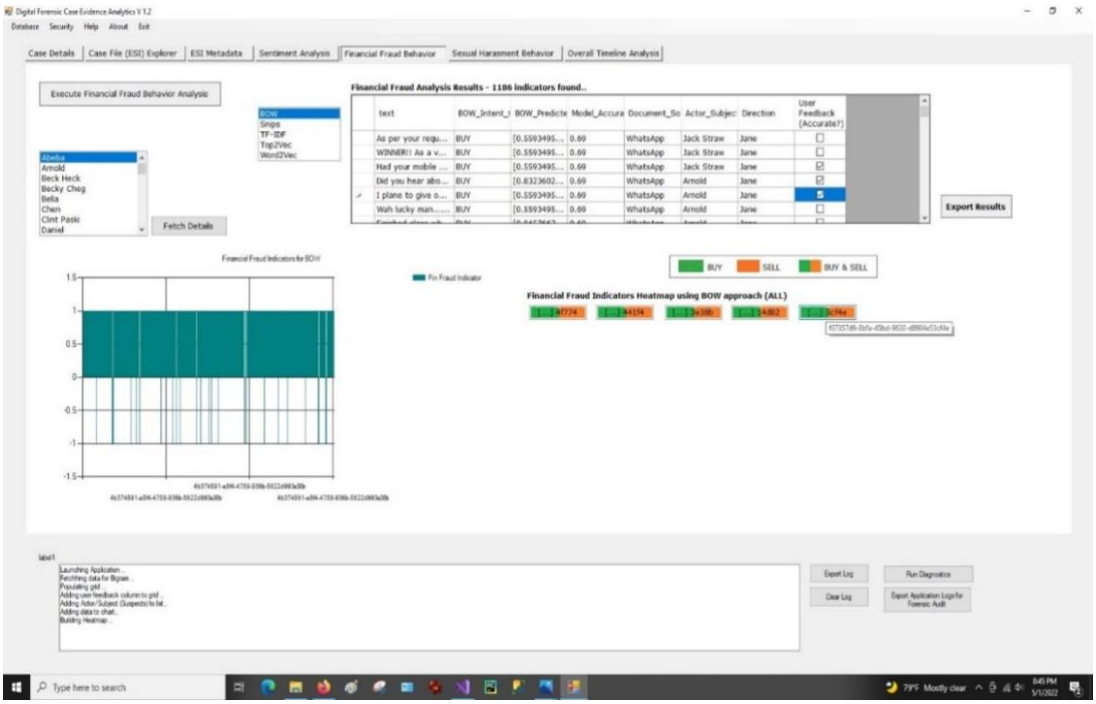

Fig. 33. Detecting financial fraud indicators using custom forensic analysis software

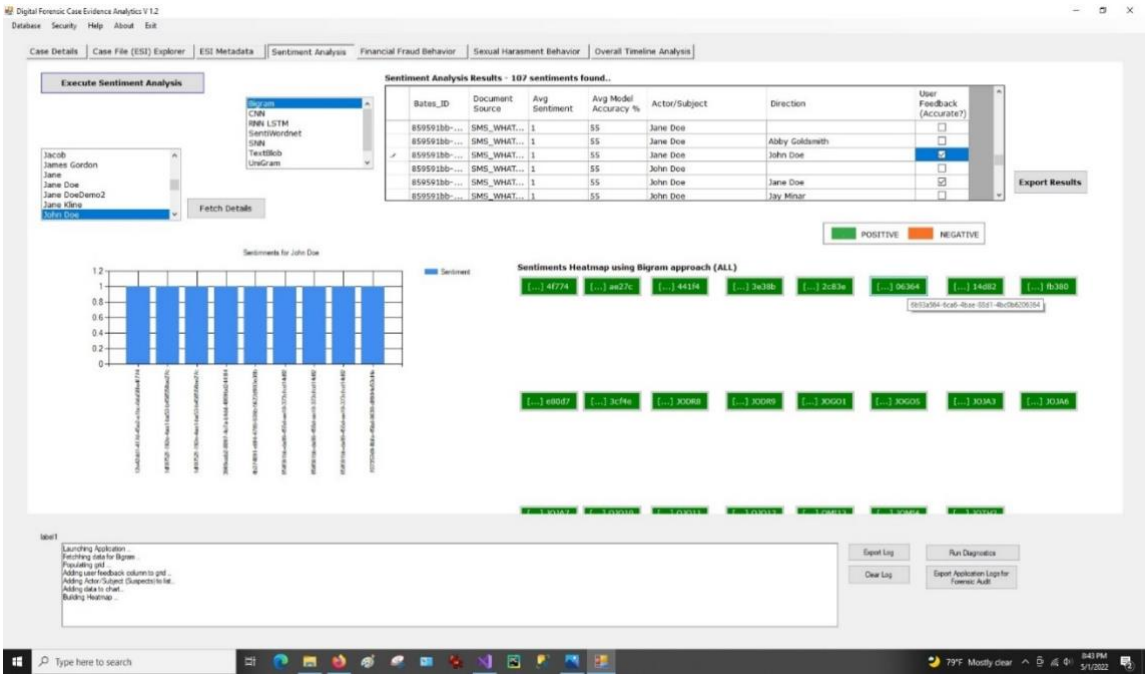

Fig. 324. Detection of Sexual Harassment evidence using the custom forensic analysis software

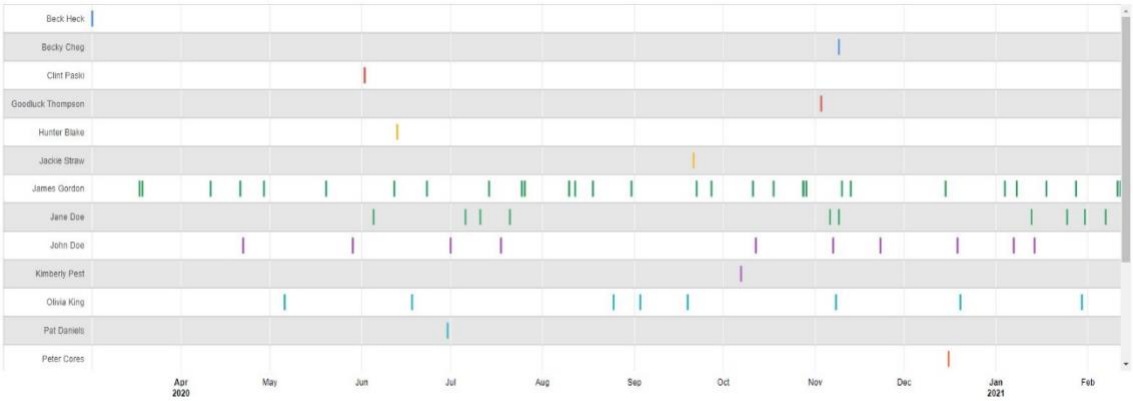

Fig. 335. Communication timelines of case suspects using Google API using the custom forensic analysis software.

Chapter VIII

CONCLUSION, LIMITATIONS AND FUTURE RESEARCH

With the eDiscovery industry and forensic investigators quickly adapting to using Machine Learning and other statistical techniques in their work, this study can significantly assist in suggesting approaches that can be leveraged to hasten the analysis of large volumes of evidence. The various statistical algorithms outlined in this study are freely available and can be immediately leveraged by investigators and eDiscovery professionals. The source code of experiments, details of approaches used, the fictitious datasets assembled, and accompanying manuscripts published on the various chapters of this study are available on GitHub for public consumption. Thus, this study has everything readily available for academic researchers and industry practitioners to implement and improve upon.

Limitations and Further Research

This study does have a few limitations. The methodology in this study is limited to the U.S. English language. However, this can be scaled into supporting other languages. The use of emoticons in today's electronic communications can convey a ton of information that can be used by criminal minds. Due to time limitations, this study skipped emoticons but accounted for emojis. Electronic communications also involve sharing of media, gifs, and images. They can be used as covert channels of communication. Due to time limitations, this study skipped such data. Risk profiling of suspect along with risk ranking techniques can be further improved. Inclusion of the interpretation of gifs in text messages, crawling of hyperlinks in text messages, and utilize social media flags such as likes and dislikes in helping flag sexual harassment indicators can be considered in future.

Focus on assessing the performance of legal analytical techniques to test and confirm the accuracy of preprocessing of evidentiary case data can also be considered for future work.

Conclusion

This dissertation aims to provide approaches and best practices when working with evidence in civil litigation. Not all evidence for the investigation or legal case is readily available. Some evidence must be forensically extracted from electronic devices, while the rest can be sourced from various devices and information technology infrastructure. All evidence together should be within the scope of the investigation or legal case. Analyzing this volume of evidence can be costly in terms of human labor and time. Thus, leveraging automation and analytics such as Machine Learning and Neural Networks can speed up evidence data analysis and greatly help locating nuggets of key evidence and their sources/origins for winning legal arguments. When analytic techniques and models are designed as suggested in this dissertation, they provide an umbrella of sub-approaches for the investigator or eDiscovery professional to choose from the best-performing one. In the case of various analytical models used in the approaches, a point to note is that the sequence and order of information from natural language is crucial for NLP modeling. Deep learning-based architectures (DNN, RNN, or attention) and Neural Network based models (CNN, LSTM) tend to model the sequence nature of natural language better than n-gram based methods (TextBlob, VADER) and lexicon-based models such as SentiWordNet and WordNet. The custom software developed in this dissertation also offers reports, heatmaps, and graphs that can then be presented during legal arguments. This dissertation also goes further into outlining areas of opportunities, challenges, and possible errors that the case investigators or eDiscovery professionals may encounter when leveraging analytics in their

experiments with digital evidence or ESI. The dissertation concludes with the development of custom software that can be referenced as a prototype or blueprint for investigators or eDiscovery professionals when they design and build their own automation interfaces or software to drive their analytical experiments or projects.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, “The Digitization of the World,” 2018. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

[2] “Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper - Cisco.” [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html#_Toc529314178. [Accessed: 31-Jan-2020].

[3] *Zubulake v. UBS Warburg LLC, 217 F.R.D. 309 (S.D.N.Y. 2003)*. .

[4] The Committee on The Judiciary House of Representatives, *Federal Rules of Civil Procedure Printed for the Use of The Committee on The Judiciary House of Representatives*. 2019.

[5] “The Basics: What is e-Discovery? | Complete Discovery Source.” [Online]. Available: https://cdslegal.com/knowledge/the-basics-what-is-e-discovery/. [Accessed: 31-Jan-2020].

[6] Robinson Rob, “An eDiscovery Market Size Mashup: 2019-2024 Worldwide Software and Services Overview,” *ComplexDiscovery*, 2019. [Online]. Available: https://complexdiscovery.com/an-ediscovery-market-size-mashup-2019-2024-worldwide-software-and-services-overview/. [Accessed: 01-Feb-2020].

[7] D. EDRM, “Processing Guide.” [Online]. Available: http://www.edrm.net/frameworks-and-standards/edrm-model/processing/.

[8] “Document Review - The Basics of E-Discovery Guide - Exterro.” [Online].

Available: https://www.exterro.com/basics-of-e-discovery/document-review-analysis-production/. [Accessed: 31-Jan-2020].

[9] O. Corporation, “What is Big Data?”

[10] B. Ingram, “Controlling E-Discovery Costs in a Big Data World,” *LexisNexis*, 2013.

[11] EDRM, “Technology Assisted Review.” [Online]. Available: https://edrm.net/resources/frameworks-and-standards/technology-assisted-review/. [Accessed: 09-Feb-2021].

[12] S. Kernisan, “TAR 1.0 or TAR 2.0: Which method is best for you?,” *Casepoint*. [Online]. Available: https://www.casepoint.com/blog/tar-1-0-versus-tar-2-0/. [Accessed: 10-Feb-2021].

[13] G. Taranto, “The Evolution of TAR,” *Law.com*, 2020. [Online]. Available: https://www.law.com/2020/12/31/the-evolution-of-tar/?slreturn=20210110063112. [Accessed: 10-Feb-2021].

[14] Casetext, *Moore v. Groupe, 868 F. Supp. 2d 137*. 2012.

[15] Justia, *Hyles v. City of New York et al, No. 1:2010cv03119 - Document 97 (S.D.N.Y. 2016)*. 2016.

[16] “What is Legal Analytics?,” *LexisNexis*, 2019. [Online]. Available: https://www.lexisnexis.com/community/lexis-legal-advantage/b/insights/posts/what-is-legal-analytics. [Accessed: 11-Feb-2021].

[17] Merriam-Webster, “Definition of Sentiment by Merriam-Webster.” [Online]. Available: https://www.merriam-webster.com/dictionary/sentiment#synonym-discussion. [Accessed: 03-Aug-2021].

[18] Dictionary.com, “Sentiment Definition & Meaning | Dictionary.com.” [Online]. Available: https://www.dictionary.com/browse/sentiment. [Accessed: 03-Aug-2021].

[19] *Rule 404 - Character Evidence; Crimes or Other Acts | 2020 Federal Rules of Evidence*. .

[20] “Litigation Forecast 2018, What Corporate Counsel Need To Know For The Coming Year,” *Crowell Moring*, 2018.

[21] “Technology Assisted Review | Predictive Coding | eDiscovery Software | Relativity.” [Online]. Available: https://www.relativity.com/ediscovery-software/technology-assisted-review/. [Accessed: 01-Feb-2020].

[22] “FRONTEO Global Website.” [Online]. Available: https://www.fronteo.com/global/kibit/. [Accessed: 01-Feb-2020].

[23] Z. Technologies, “EDRM and ZL Launch New Enron Email Data Set – ZL Tech.” [Online]. Available: https://www.zlti.com/press-releases/edrm-and-zl-launch-new-enron-email-data-set/. [Accessed: 25-Oct-2021].

[24] “U.S. District Courts—Civil Cases Commenced, by Nature of Suit,During the 12-Month Periods Ending September 30, 2014 through 2018.”

[25] “U.S. District Courts—Civil Cases Commenced, by Basis of Jurisdiction and Nature of Suit,During the 12-Month Periods Ending September 30, 2018 and 2019.”

[26] KPMG Australia, “Global Banking Fraud Survey 2019,” Jun. 2019.

[27] “EDRM Model | EDRM.” [Online]. Available: https://www.edrm.net/resources/frameworks-and-standards/edrm-model/.

[Accessed: 31-Jan-2020].

[28] "Searchable e-Discovery Case Log," *K&L Gates*. [Online]. Available: https://ediscovery.klgates.com/. [Accessed: 10-Jul-2020].

[29] Logikcull, "2019 eDiscovery Billing & Cost Recovery Survey," *Logikcull*, 2020. [Online]. Available: https://www.logikcull.com/public/files/Logikcull-2019-eDiscovery-Billing-Survey.pdf.

[30] N. Pace and L. Zakaras, *Where the money goes: Understanding litigant expenditures for producing electronic discovery*. 2012.

[31] D. Law, "EDRM Model," *Duke Law*. [Online]. Available: https://www.edrm.net/frameworks-and-standards/edrm-model/. [Accessed: 01-Jan-2021].

[32] "Legal (Litigation) Holds and Data Preservation - Basics… | Logikcull." [Online]. Available: https://www.logikcull.com/guide/legal-holds-data-preservation. [Accessed: 08-Jul-2020].

[33] "Current Listing of States That Have Enacted E-Discovery Rules | Electronic Discovery Law," *K&L Gates, Electronic Discovery Law*. [Online]. Available: https://www.ediscoverylaw.com/state-district-court-rules/. [Accessed: 10-Jul-2020].

[34] C. Delgado, "Facing E-Discovery: Preparedness and Compliance," *Georgia State University College of Law*, 2011. [Online]. Available: https://www.semanticscholar.org/paper/Facing-E-Discovery%3A-Preparedness-and-Compliance-Delgado/bc271aef8ca642cf5bd27d945d71b51f9cca948b. [Accessed: 10-Jul-2020].

[35] “Information Governance Reference Model | EDRM.” [Online]. Available: https://www.edrm.net/resources/frameworks-and-standards/information-governance-reference-model/. [Accessed: 08-Jul-2020].

[36] E. M. Negangard and R. G. Fay, “Electronic discovery (Ediscovery): Performing the early stages of the enron investigation,” *Issues Account. Educ.*, vol. 35, no. 1, pp. 43–58, Feb. 2020.

[37] “The Enron Data Sets Cleansed by Nuix and EDRM,” *Nuix and EDRM*. [Online]. Available: https://info.nuix.com/Enron.html. [Accessed: 01-Feb-2020].

[38] R. J. Bernier, “Avoiding an E-Discovery Odyssey.” 01-Jul-2009.

[39] H. Hyman and F. I. Warren, “Using Bag of Words (BOW) and Standard Deviations to Represent Expected Structures for Document Retrieval,” in *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA*, 2010.

[40] H. Hyman, T. Sincich, R. Will, M. Agrawal, B. Padmanabhan, and W. Fridy, “A process model for information retrieval context learning and knowledge discovery,” *Artif. Intell. Law*, vol. 23, no. 2, pp. 103–132, Jun. 2015.

[41] A. Hernandez, “Common Problems With E-Discovery—and Their Solutions,” *The Federal Lawyer*, 2016.

[42] “New EDRM Enron Email Data Set | EDRM,” *EDRM*. [Online]. Available: https://www.edrm.net/resources/data-sets/edrm-enron-email-data-set/. [Accessed: 01-Feb-2020].

[43] D. Austin, “The Enron Data Set is No Longer a Representative Test Data Set: eDiscovery Best Practices,” *CloudNine Blog*. [Online]. Available:

https://cloudnine.com/ediscoverydaily/electronic-discovery/the-enron-data-set-is-no-longer-a-representative-test-data-set-ediscovery-best-practices/. [Accessed: 03-May-2020].

[44] D. Noever, “The Enron Corpus: Where the Email Bodies are Buried?,” Jan. 2020.

[45] C. Faklaris and S. A. Hook, “‘Oh, Snap! The State of E-Discovery as Social Media Goes Mobile via Snapchat, WhatsApp and Other Messaging Apps,’” Office of the Vice Chancellor for Research, Apr. 2003.

[46] R. A. Calix and G. M. Knapp, “Actor level emotion magnitude prediction in text and speech,” *Multimed. Tools Appl.*, vol. 63, no. 3, pp. 319–332, Oct. 2013.

[47] G. Tripathi, K. Singh, and D. K. Vishwakarma, “Convolutional neural networks for crowd behaviour analysis: a survey,” *Vis. Comput.*, vol. 35, no. 5, pp. 753–776, May 2019.

[48] E. Romera, L. M. Bergasa, and R. Arroyo, “Need data for driver behaviour analysis? Presenting the public UAH-DriveSet,” in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2016, pp. 387–392.

[49] S. Garg, A. K. Singh, A. K. Sarje, and S. K. Peddoju, “Behaviour analysis of machine learning algorithms for detecting P2P botnets,” in *2013 15th International Conference on Advanced Computing Technologies, ICACT 2013*, 2013.

[50] F. Haddadi, J. Morgan, E. G. Filho, and A. N. Zincir-Heywood, “Botnet behaviour analysis using IP flows: With http filters using classifiers,” in *Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014*, 2014, pp. 7–12.

[51] Shalini and D. Singh, “Comparative Analysis of Clustering Techniques for

Customer Behaviour," in *Advances in Intelligent Systems and Computing*, 2018, vol. 584, pp. 753–763.

[52] R. M. Foxx, "Applied Behavior Analysis Treatment of Autism: The State of the Art," *Child and Adolescent Psychiatric Clinics of North America*, vol. 17, no. 4. pp. 821–834, Oct-2008.

[53] B. Liu, "Opinion Mining and Sentiment Analysis," in *Web Data Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 459–526.

[54] Z. Xiaomei, Y. Jing, Z. Jianpei, and H. Hongyu, "Microblog sentiment analysis with weak dependency connections," *Knowledge-Based Syst.*, vol. 142, pp. 170–180, Feb. 2018.

[55] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," *Math. Comput. Appl.*, vol. 23, no. 1, p. 11, Feb. 2018.

[56] "Behavior Analysis - Association for Behavior Analysis International." [Online]. Available: https://www.abainternational.org/about-us/behavior-analysis.aspx. [Accessed: 19-Jul-2020].

[57] "Behavioral Analysts — FBI." [Online]. Available: https://www.fbi.gov/audio-repository/news-podcasts-inside-bau-profilers.mp3/view. [Accessed: 19-Jul-2020].

[58] "National Board of Forensic Evaluators, Inc. - CFBA." [Online]. Available: https://www.nbfe.net/CFBA1. [Accessed: 19-Jul-2020].

[59] D. Coderre, "Computer-Assisted Techniques for Fraud Detection," *CPA J.*, vol. 69, no. 8, 1999.

[60] Z. Rezaee, "Causes, consequences, and deterence of financial statement fraud," *Z.*

Rezaee / Crit. Perspect. Account., vol. 16, pp. 277–298, 2005.

[61] P. Dunn, “The Impact of Insider Power on Fraudulent Financial Reporting,” *J. Manage.*, vol. 30, no. 3, pp. 397–412, Jun. 2004.

[62] R. Davidson, A. Dey, and A. Smith, “Executives’ ‘off-the-job’ behavior, corporate culture, and financial reporting risk,” *J. financ. econ.*, vol. 117, no. 1, pp. 5–28, Jul. 2015.

[63] M. Weatherford, “Mining for Fraud,” *IEEE Intell. Syst.*, vol. 17, no. 4, pp. 4–6, 2002.

[64] P. Richhariya and P. Singh, “A survey on financial fraud detection methodologies,” *Int. J. Comput. Appl.*, vol. 45, no. 22, 2012.

[65] S. Ghosh and D. L. Reilly, “Credit card fraud detection with a neural-network,” in *Proceedings of the Hawaii International Conference on System Sciences*, 1994, vol. 3, pp. 621–630.

[66] M. Syeda, Y. Q. Zhang, and Y. Pan, “Parallel granular neural networks for fast credit card fraud detection,” in *IEEE International Conference on Fuzzy Systems*, 2002, vol. 1, pp. 572–577.

[67] E. L. Barse, H. Kvarnström, and E. Jonsson, “Synthesizing test data for fraud detection systems,” in *Proceedings - Annual Computer Security Applications Conference, ACSAC*, 2003, vol. 2003-Janua, pp. 384–394.

[68] C. C. Chiu and C. Y. Tsai, “A web services-based collaborative scheme for credit card fraud detection,” in *Proceedings - 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service, EEE 2004*, 2004, pp. 177–181.

[69] A. Deshmukh and L. Talluru, “A rule-based fuzzy reasoning system for assessing

the risk of management fraud," *Intell. Syst. Accounting, Financ. Manag.*, vol. 7, no. 4, pp. 223–241, Dec. 1998.

[70] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intell. Syst. Their Appl.*, vol. 14, no. 6, pp. 67–74, 1999.

[71] D. Choi and K. Lee, "An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation," *Secur. Commun. Networks*, vol. 2018, 2018.

[72] R. R. Vanasco, "Fraud auditing," *Manag. Audit. J.*, vol. 13, no. 1, pp. 4–71, Feb. 1998.

[73] S. Ramamoorti, D. E. Morrison, J. W. Koletar, and K. R. Pope, *A.B.C.'s of Behavioral Forensics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[74] M. A. Paludi and R. B. Barickman, *Academic and Workplace Sexual Harassment: A Resource Manual*. SUNY Press (November 15, 1991), 1991.

[75] W. B. Dziech and L. Weiner, *The Lecherous Professor: Sexual Harassment on Campus*. 1990.

[76] "Sexual Harassment | U.S. Equal Employment Opportunity Commission." [Online]. Available: https://www.eeoc.gov/sexual-harassment. [Accessed: 06-Apr-2022].

[77] *Title VII, Civil Rights Act of 1964, as amended | U.S. Department of Labor*. .

[78] I. Rodríguez-Rodríguez and P. Heras-González, "How are universities using Information and Communication Technologies to face sexual harassment and how can they improve?," *Technol. Soc.*, vol. 62, p. 101274, Aug. 2020.

[79] T. Bauer, E. Devrim, M. Glazunov, W. L. Jaramillo, B. Mohan, and G. Spanakis, "#MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment," *Commun. Comput. Inf. Sci.*, vol. 1167 CCIS, pp. 503–521, 2020.

[80] A. Garrett and N. Hassan, "Understanding the silence of sexual harassment victims through the #Whyididntreport movement," *Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2019*, pp. 649–652, Aug. 2019.

[81] M. Saeidi, S. B. Samuel, E. Milios, N. Zeh, and L. Berton, "Categorizing Online Harassment on Twitter," *Commun. Comput. Inf. Sci.*, vol. 1168 CCIS, pp. 283–297, 2020.

[82] E. Alawneh, M. Al-Fawa'Reh, M. T. Jafar, and M. Al Fayoumi, "Sentiment analysis-based sexual harassment detection using machine learning techniques," *Proceeding - 2021 Int. Symp. Electron. Smart Devices Intell. Syst. Present Futur. Challenges, ISESD 2021*, Jun. 2021.

[83] P. Basu, T. Singha Roy, S. Tiwari, and S. Mehta, "CyberPolice: Classification of Cyber Sexual Harassment," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12981 LNAI, pp. 701–714, 2021.

[84] M. Sathiyanarayanan, "Improving visual investigation analysis of digital communication data within e-discovery," 2017.

[85] A. Jarrett and K.-K. R. Choo, "The impact of automation and artificial intelligence on digital forensics," *Wiley Interdiscip. Rev. Forensic Sci.*, vol. 3, no. 6, p. e1418, Nov. 2021.

[86] F. Mitchell, "The use of Artificial Intelligence in digital forensics: An introduction

- SAS-Space," *Digit. Evid. Electron. Signat. Law Rev.*, vol. 7, 2010.

[87] P. H. Rughani, "Artificial Intelligence Based Digital Forensics Framework.," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 8, 2017.

[88] I. Baggili and V. Behzadan, "Founding The Domain of AI Forensics," *CEUR Workshop Proc.*, vol. 2560, pp. 31–35, Dec. 2019.

[89] P. Bhatt, "Machine Learning Forensics: A New Branch Of Digital Forensics," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 8, pp. 217–222, Aug. 2017.

[90] "Study Github Repository." .

[91] "Speech error - Wikipedia," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Speech_error. [Accessed: 10-Feb-2021].

[92] S. Krishnan and N. Shashidhar, "Interplay of Digital Forensics in eDiscovery," *Int. J. Comput. Sci. Secur.*, vol. 15, no. 2, p. 19, 2021.

[93] J. Greer, "Email Threading in eDiscovery: The Longest Thread Policy," *Digital WarRoom*, 2019. [Online]. Available: https://www.digitalwarroom.com/blog/email-threading-ediscovery-problems-with-longest-thread. [Accessed: 10-Feb-2021].

[94] "Definition: Model fitting," *Edpresso*. [Online]. Available: https://www.educative.io/edpresso/definition-model-fitting. [Accessed: 11-Feb-2021].

[95] NIST, "Guidelines for Media Sanitization, Special Publication 800-88." 2014.

[96] N. I. S. Program, *DoD 5220.22-M, Operating Manual*. 2006.

[97] Z. Technologies, "Hillary Clinton's Emails | Kaggle." [Online]. Available: https://www.kaggle.com/kaggle/hillary-clinton-emails?select=Emails.csv.

[Accessed: 25-Oct-2021].

[98] "UCI Machine Learning Repository: SMS Spam Collection Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection#. [Accessed: 25-Oct-2021].

[99] "Twitter US Airline Sentiment," *Kaggle*, 2019. [Online]. Available: https://www.kaggle.com/crowdflower/twitter-airline-sentiment/activity. [Accessed: 14-Feb-2021].

[100] R. Agarwal, "Twitter hate speech," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/vkrahul/twitter-hate-speech. [Accessed: 14-Feb-2021].

[101] "First GOP Debate Twitter Sentiment," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/crowdflower/first-gop-debate-twitter-sentiment. [Accessed: 14-Feb-2021].

[102] J. Bencina, "Facebook News Scraper," *Github*, 2017. [Online]. Available: https://github.com/jbencina/facebook-news. [Accessed: 14-Feb-2021].

[103] K. Zhang *et al.*, "SES: Sentiment elicitation system for social media data," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2011, pp. 129–136.

[104] "Bag of Words Meets Bags of Popcorn, Use Google's Word2Vec for movie reviews," *Kaggle*, 2014. [Online]. Available: https://www.kaggle.com/c/word2vec-nlp-tutorial/data. [Accessed: 25-Oct-2021].

[105] "Pstxy: Outlook .pst and .ost file reader for .Net," *pantilesoft, Github*. [Online]. Available: https://github.com/pantilesoft/pantilesoft.github.io. [Accessed: 14-Feb-2021].

[106] C. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 8, no. 1, pp. 216–225, May 2014.

[107] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.

[108] “WordNet | A Lexical Database for English,” *Princeton University*, 2010. [Online]. Available: https://wordnet.princeton.edu/. [Accessed: 10-Oct-2021].

[109] “TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation.” [Online]. Available: https://textblob.readthedocs.io/en/dev/. [Accessed: 10-Oct-2021].

[110] J. Pennington, R. Socher, and C. Manning, “{G}lo{V}e: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, 2014, pp. 1532–1543.

[111] D. Kaihua, “A simple deep neural network that beats TextBlob and VADER packages at sentiment classification,” *Medium*, 2021. [Online]. Available: https://medium.com/@dingkaihua/a-simple-deep-neural-network-that-beats-textblob-and-vader-packages-at-sentiment-classification-bd4990b8cf3b. [Accessed: 10-Jul-2022].

[112] S. Krishnan, “Project · GitHub.” [Online]. Available: https://github.com/kshsus. [Accessed: 06-May-2022].

[113] S. Krishnan, N. Shashidhar, C. Varol, and A. Rezbaul Islam, “Evidence Data Preprocessing for Forensic and Legal Analytics,” *Int. J. Comput. Linguist.*, vol. 12,

no. 2, pp. 24–34, 2021.

[114] “reddit.com: api documentation,” *reddit.com*. [Online]. Available: https://www.reddit.com/dev/api/. [Accessed: 04-Feb-2022].

[115] R. Aroussi, “yfinance · PyPI,” *Apache Software License (Apache)*. [Online]. Available: https://pypi.org/project/yfinance/. [Accessed: 04-Feb-2022].

[116] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018.

[117] D. Angelov, “Top2Vec: Distributed Representations of Topics,” 2020.

[118] A. Coucke *et al.*, “Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces,” May 2018.

[119] “Snips Natural Language Understanding — Snips NLU 0.20.2 documentation.” [Online]. Available: https://snips-nlu.readthedocs.io/en/latest/. [Accessed: 05-Feb-2022].

[120] “Open source conversational AI,” *Rasa*. [Online]. Available: https://rasa.com/. [Accessed: 06-Feb-2022].

[121] “NLU-benchmark/2017-06-custom-intent-engines at master · sonos/nlu-benchmark,” *GitHub*. [Online]. Available: https://github.com/sonos/nlu-benchmark/tree/master/2017-06-custom-intent-engines. [Accessed: 05-Feb-2022].

[122] GitHub, “nlu-benchmark,” *GitHub*. [Online]. Available: https://github.com/wenjingu/nlu-benchmark. [Accessed: 05-Feb-2022].

[123] M. Terblanche and V. Marivate, “Loughran McDonald-SA-2020 Sentiment Word

List," 2021. [Online]. Available: https://researchdata.up.ac.za/articles/dataset/Loughran_McDonald-SA-2020_Sentiment_Word_List/14401178. [Accessed: 06-Feb-2022].

[124] M. Britney, "BERT 101 - State Of The Art NLP Model Explained," *Hugging Face,* 2022. [Online]. Available: https://huggingface.co/blog/bert-101. [Accessed: 10-Jul-2022].

[125] H. Mclaughlin, C. Uggen, and A. Blackstone, "Sexual Harassment, Workplace Authority, and the Paradox of Power," *Am. Sociol. Rev.*, vol. 77, no. 4, pp. 625–647.

[126] "Power's Role In Sexual Harassment - WSJ," *Wall Street Journal,* 2018. [Online]. Available: https://www.wsj.com/articles/powers-role-in-sexual-harassment-1517844769. [Accessed: 15-Apr-2022].

[127] "The Psychological Persuasion Techniques of Sexual Predators | Psychology Today." [Online]. Available: https://www.psychologytoday.com/us/blog/the-new-teen-age/201905/the-psychological-persuasion-techniques-sexual-predators. [Accessed: 15-Apr-2022].

[128] F. F. Nova, R. Rifat, P. Saha, S. I. Ahmed, and S. Guha, "Online sexual harassment over anonymous social media in Bangladesh," *ACM Int. Conf. Proceeding Ser.*, Jan. 2019.

[129] D. K. Gyawali, "Sexual Harassment and Its effects on Mental Health OF THE Teenage School Girls in Lalitpur and rupandehi district," *J. Balkumari Coll.*, vol. 10, no. 1, pp. 39–47, Jan. 2021.

[130] W. Crebbin *et al.*, "Prevalence of bullying, discrimination and sexual harassment

in surgery in Australasia," *ANZ J. Surg.*, vol. 85, no. 12, pp. 905–909, Dec. 2015.

[131] Sundar Krishnan, N. Shashidhar, C. Varol, and A. R. Islam, "Evidence Data Preprocessing for Forensic and Legal Analytics," *Int. J. Comput. Linguist.*, vol. 12, no. 2, pp. 24–34, 2021.

[132] Nicapotato, "Women's E-Commerce Clothing Reviews," *Kaggle*, 2018. [Online]. Available: https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews. [Accessed: 15-Apr-2022].

[133] A. Cercas Curry, G. Abercrombie, and V. Rieser, "{C}onv{A}buse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational {AI}," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7388–7403.

[134] M. Rezvan, K. Thirunarayan, S. Shekarpour, V. L. Shalin, L. Balasuriya, and A. Sheth, "A quality type-aware annotated corpus and lexicon for harassment research," *WebSci 2018 - Proc. 10th ACM Conf. Web Sci.*, pp. 33–36, May 2018.

[135] A. Coucke *et al.*, "Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces," May 2018.

[136] "NLP vs. NLU: What's the Difference and Why Does it Matter?," *Rasa, The Rasa Blog*. [Online]. Available: https://rasa.com/blog/nlp-vs-nlu-whats-the-difference/. [Accessed: 15-Feb-2022].

[137] S. C. of I. Nisha Priya Bhatia v. Union of India & Anr. CA No. 2365/2020, "Types of Sexual Harassment." .

[138] M. Herman *et al.*, "NIST Cloud Computing Forensic Science Challenges."

[139] L. Pan and L. Batten, "Digital Forensic Research Conference Reproducibility of

Digital Evidence in Forensic Investigations," 2005.

[140] "Randomization | Data Preparation and Feature Engineering for Machine Learning | Google Developers." [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/randomization. [Accessed: 02-May-2022].

[141] "Tour of Data Sampling Methods for Imbalanced Classification." [Online]. Available: https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/. [Accessed: 02-May-2022].

[142] D. Carty, "Training, Validation and Testing Data Explained," *Applause, Blog / Dev & QA Trends / TTraining Data vs. Validation Data vs. Test Data for ML Algorithms*, 2021. [Online]. Available: https://www.applause.com/blog/training-data-validation-data-vs-test-data. [Accessed: 04-May-2022].

[143] S. W. Hall, | Amin Sakzad, | Kim-Kwang, and R. Choo, "Explainable artificial intelligence for digital forensics," *Wiley Interdiscip. Rev. Forensic Sci.*, vol. 4, no. 2, p. e1434, Mar. 2022.

[144] K. Reeve, "Python to Pseudocode converter," *BlueNexus , GitHub*. [Online]. Available: https://gist.github.com/BlueNexus/599962d03a1b52a8d5f595dabd51dc34.

APPENDIX

Pseudo code, Program Code and ESI Datasets

Python code to pseudo code conversion was performed using a software utility - gists [144]. The raw program code and pseudo code for various Python, the fictional datasets used, Visual Studio project files and C# programs are also available on GitHub [112].

Relevant Dissertation Papers

Published:

1. *Evidence Data Preprocessing for Forensic and Legal Analytics*, S Krishnan, N Shashidhar, C Varol, A Islam, International Journal of Computational Linguistics (IJCL)

Pending publication in August/2022

1. *Sentiment Analysis of Case Suspects in Digital Forensics and Legal Analytics*, S Krishnan, N Shashidhar, C Varol, A Islam - International Journal of Security (IJS).
2. *A Novel Text Mining Approach to Securities and Financial Fraud Detection of Case Suspects*, S Krishnan, N Shashidhar, C Varol, A Islam - International Journal of Artificial Intelligence and Expert Systems (IJAE).
3. *A Novel Text Mining Approach to Sexual Harassment Detection of Case Suspects*, S Krishnan, N Shashidhar, C Varol, A Islam - International Journal of Artificial Intelligence and Expert Systems (IJAE).
4. *Analytics in Digital Forensics and eDiscovery Software - DevOps, Opportunities and Challenges,* S Krishnan, N Shashidhar, C Varol, A Islam - International Journal of Security (IJS)

VITA

SUNDAR KRISHNAN

Possess a career experience of 26 years in Information technology that includes about 10 years in Cybersecurity. Passionate to transfer industry gained knowledge through teaching and mentoring students in addition to undertaking development of academic Cybersecurity programs involving Penetration Testing, Defense, Incident Forensics, Privacy and Risk management.

EDUCATION

- PhD Candidate in Cyber and Digital Forensics, Sam Houston State University, Expected graduation August/2022
- Master's in Digital Forensics, SHSU, May/2015
- Master's in Computer Applications, Bharathiar University, India, Dec/2002
- Bachelor of Science (Major in Electronics) Bangalore University, India, May/1995

CERTIFICATIONS

- Certified Information Systems Security Professional (CISSP)
- Certified Data Privacy Solutions Engineer (CDPSE)
- Project Management Professional (PMP)
- Six Sigma Black Belt
- ITIL v3 Foundation•Certified Information Security Manager (CISM)
- SEI-CMMI ATM (Assessment Team Member SCAMPI-B)
- Microsoft Certified Professional in 70-562 (MCP .NET 3.5 and Web application)
- Certified in Industrial Control Systems – Cybersecurity, ICS-CERT, DHS

WORK EXPERIENCE

- Graduate Research Assistant, Sam Houston State University, Huntsville, TX, Aug 2018 - till date

Worked on research projects on Smartphone, Disk and Network Forensics. Taught classes assigned by the Department Chair. Authored and co-authored peer-review publications, and conference presentations

- Information Security Summer Intern Sam Houston State University, June/2021 - Aug/2021

Worked with the SHSU Information Security Team a part of my PhD program required Internship.

- Information Security Operations Manager, Methodist Le Bonheur Healthcare, Memphis, TN, Mar/2017 - June/2018

Responsible for managing the information security operations across the organization thereby keeping an eye on organization's digital security footprint. Managed a team of security professionals and third-party SOC resources. Responsible for the enterprise security program in the areas of risk management, vulnerability management, incident response, security operations, governance, training, compliance, and incident forensics.

- Information Systems Security Manager, University Health System, San Antonio, TX, Nov/2015 - Mar/2017

Responsible for implementing the information systems security program from ground-up across the organization. Responsible for the enterprise security program in the areas of risk

management, vulnerability management, incident response, security operations, governance, and compliance.

- Principal Consultant, Security Intelligence Analytics & Assurance, Wipro Ltd, Houston, TX, USA, Jun/2011 - Nov/2015

Worked as a Principal Consultant in the Enterprise Security Solutions (ESS) practice as part of the Security Intelligence Analytics Assurance (SIAA) team. Provided consulting services (advisory services, transformational solutions, and managed security services) around end-to-end security and compliance solutions globally across industry verticals.

- Cybersecurity Summer Intern, InduSoft, Invensys (Schneider Electric), Austin, TX, May/2014 - Aug/2014

Created documents and presentations, collaborated with the QA team on best practices in industrial cyber security, based on the products and services offered by InduSoft.

- Technical Lead/Project Manager, Science Applications International Corporation (SAIC), Houston, TX, Sept/2003 - Jun/2011

Led and managed full development lifecycle projects of various sizes and complexities. Coordinated the timeliness and quality of work effort for the team onshore and offshore. Managed a team of five resources based in Houston, TX, one resource based in UK and ten resources based out of India. Managed escalations for project related issues, challenges, or questions from project team. Played a role of Lead architect and Operations Coordinator (technical) with Oil and Gas client, involving 250 applications.

- Worked as Software Developer/Programmer at various Information Technologies companies in India between June’ 1995 to Sept’ 2003