Executive Summary

This report seeks to present and critically evaluate the application of Artificial Intelligence and Machine Learning within Electronic Discovery and Digital Forensics. The objectives of this report are to:

- present summaries of relevant literature on machine learning in eDiscovery
- critically analyse the relevance of machine learning in eDiscovery
- compare machine learning protocols.

Methodology and Literature Review

To understand the usefulness of Data Analytics and Machine Learning in the eDiscovery process as well as in the field of Cybersecurity, the Electronic Discovery Reference Model was explored to obtain an understanding of the eDiscovery process, identifying constraints within the process in practice and how Data Analytics and Machine Learning makes the process much more efficient. The underlying algorithms leveraged within the Electronic Discovery Reference Model as well as their applications were also explored to understand how they are related to the Electronic Discovery Reference Model and different use cases for each application. Version 2.0 of the Good Practice Discovery Guide by the Commercial Litigation Association of Ireland was also consulted to understand best practices when performing eDiscovery.

EDRM.net defines Electronic Discovery (eDiscovery) as “the exchange, review and analysis of Electronically Stored Information (ESI) with the goal of bringing to light relevant information for a trial, arbitration or a hearing”. Electronic Discovery Reference Model (EDRM) is a conceptual model, representing the different phases of the eDiscovery process from identification of relevant material to a case through to the presentation of the material in a court of law. The EDRM shows the steps through which ESI moves from its original location to a trial or hearing as well as the framework used to govern ESI, all done iteratively. (EDRM, n.d.)

Fig. 1 is a diagram that shows the phases of eDiscovery:

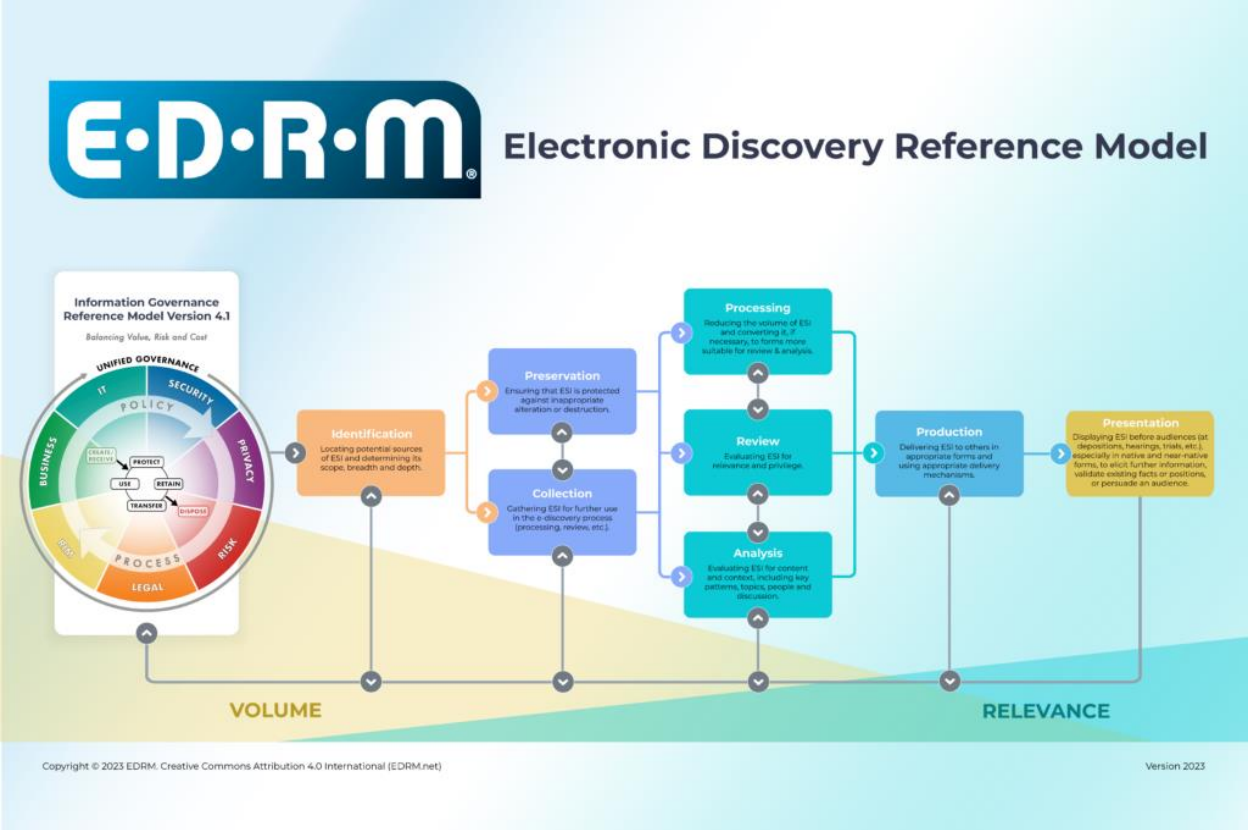

Fig.1 Detailed version of the Electronic Discovery Reference Model (EDRM*, 2023)*

Technology Assisted Review: is the use of computing technology to reduce the risk and cost associated of a review of documents in electronic discovery. It is split into two broad categories namely:

- Analytics: technology arranges documents in a way that make it easier to perform a traditional review. These technologies are usually used in the earlier phases of eDiscovery to prioritize relevant documents for further review. (CLAI, 2015)Techniques used in this category include keyword searching, deduplication and email threading.
- Predictive Coding: technology enables decision-making by algorithms trained by expert reviewers to identify relevant, responsive and privileged documents. (CLAI, 2015)This category of technology employs machine learning techniques to identify documents based on a seed set of example documents correctly labelled by the human reviewer.

Machine learning is the use of data and algorithms to increase the accuracy of software programs performing tasks. The two main approaches used to train machine learning algorithms are supervised and unsupervised learning. (IBM, n.d.-a)

Supervised learning: This approach trains algorithms using input and output data that is already correctly identified to predict outcomes whenever data is fed into it. (IBM, n.d.-b) The outcomes are predicted by classification which separates data based on sample input and output fed into

it and determines the relationship that exists between the known and unknown variables. These algorithms are usually present at the earlier stages of Processing and Review of ESI to identify keywords, duplicate or near-duplicate documents and email threads for further review. (CLAI, 2015)

Unsupervised learning: This approach trains machine learning algorithms by feeding unlabeled data to the algorithm to find relationships between the data based on a set of rules. Outcomes are predicted by an algorithm clustering similar data. (IBM, n.d.-c)

Natural language processing: This is the part of machine learning that deals with training algorithms to understand, analyse and generate written and spoken language that humans use to communicate with each other. (P. Ghavami, 2019) These natural language processing algorithms parse data from the ESI that has been deemed relevant to the case for analysis either by the reviewers or other machine learning programs.

Keyword searching: This is the review of documents in search of the presence of keywords compiled by the legal practitioners to identify documents that may be relevant to the case. (CLAI, 2015) Keyword searches on documents are usually carried out using Boolean operators to search for certain keywords occurring in the same document and wildcards to find closely matching words or phrases to the keyword. Text tokenization and normalization is also done for keyword searching to ensure text can be easily searched based on the occurrence of individual words or phrases in documents.

Deduplication: Due to the volume of ESI and similarities of different documents, exact duplicate and near duplicate texts must be appropriately identified to properly cull the amount of ESI for review. (Pace & Zakaras, 2012) Deduplication is usually performed at the Processing stage using techniques such as Locality Sensitive Hashing to give documents a unique digital fingerprint by calculating the hash and the algorithm compares the fingerprints of documents to find exact or near duplicates.

Simple Active Learning: This is a supervised Analysis of ESI by an algorithm. The algorithm is given a set of seed documents that are selected randomly or using keywords and subsequent training documents fed to the algorithm are ones the algorithm is least certain about. (Cormack & Grossman, 2014)

Continuous Active Learning: This is end-to-end Review and Analysis of ESI by an algorithm. The algorithm is given a set of rules and chooses a seed set of documents based on the rules and proceeds to analyse the documents, clustering the documents with the most similarity together. (Cormack & Grossman, 2014) This is an unsupervised learning approach as the algorithm is not trained prior to the processing of requests. Documents are fed into the algorithm after the Processing stage where documents with relevant keywords have been identified for review and analysis and this is done iteratively.

Findings

Relevance and use of machine learning in eDiscovery

Machine learning has been used to increase the speed and accuracy of electronic discovery by utilizing algorithms to quickly identify, process and review documents based on certain rules, saving time and cost that would have been expended if a manual review was done. With the prevalence of a large volume of ESI within discovery in modern litigation and the estimated average cost of review being 73 cents per dollar spent on eDiscovery, it has become crucial to mitigate the mistakes of inconsistent human reviewers reviewing documents and find ways to lower costs.(Pace & Zakaras, 2012)

A report by the RAND Corporation on a study of a corpus of emails from the investigation of Enron to determine the effectiveness of human reviewers and predictive coding algorithms with a topic authority acting as the final reviewing authority to determine correctness of each document's classification and across three metrics. (Pace & Zakaras, 2012) The human reviewers were represented by law students and professionals while two computer applications were used. Each test was a comparison between an application's decision and a human reviewer's decision, appealed on. The three metrics are explained below:

Recall: is the measurement of relative completeness of the identification of specific items of interest to the total number of items in a set of data. (Pace & Zakaras, 2012) The applications out-performed human reviewers in three of the five topics.

Precision: is the measurement of documents retrieved from a corpus relative to the number of documents correctly identified. (Pace & Zakaras, 2012) The applications out-performed human reviewers in all five topics.

F-Measure: is a measure of the relationship between precision and recall, with larger f-measure indicating better overall results. (Pace & Zakaras, 2012) The applications out-performed humans in four of the five topics.

While it is not made abundantly clear as to how much is saved using predictive coding for eDiscovery relative to human reviewers, there is no doubt that predictive coding is more efficient based on the empirical evidence presented in the report. (Pace & Zakaras, 2012)

Table 1 below shows a comparison of the Continuous Active Learning and Simple Active Learning machine learning protocols:

Continuous Active Learning	**Simple Active Learning**
Aims to find as many responsive documents as quickly as possible	Aims to create the best classifier to identify documents for further manual review
End-to-end process as algorithm identifies seed documents based on defined rules	Dependent on human reviewer to define seed document set

Training continues until the review process is complete	Training is suspended whenever the ideal classifier is produced
All documents are consistently evaluated based on defined rules	Documents that have already been identified based on seed document set may not be reviewed further
More adaptable to changes in corpus and issues of relevance	Less adaptable to changes in corpus and issues of relevance

Table 1. Comparison of CAL and SAL

In most litigation, ESI is text-based and most algorithms process mostly text data. (CLAI, 2015) ESI in other formats is unable to be processed by predictive coding. Predictive coding is more suited to situations where there is a large volume of text-based data to determine relevance. For multimedia data such as audio or video, the data will have to be transcribed so searchable documents can be generated.

Conclusion

While technology assisted review has been proven to reduce the time needed for eDiscovery, the efficacy of the claim of machine learning to reduce costs has been explored as well with no definite answers being provided in absolute terms. The cost savings usually varies based on the unique situation that is considered. A significant limitation of ESI within the context of eDiscovery is that algorithms usually process text-based information and not multimedia. This means that manual effort will still need to be applied to process multimedia data.

Further research may explore leveraging advanced natural language processing algorithms within eDiscovery to parse audio files for review and analysis. Cost savings for specific situations may also be explored to create better leverage for decision making for legal practitioners looking to perform technology assisted review.

References

0.0 Introduction—EDRM. (n.d.). Retrieved April 18, 2024, from https://edrm.net/wiki/introduction/

Chapter 4. Natural Language Processing. (2019). In P. Ghavami, *Big Data Analytics Methods* (pp. 65–84). De Gruyter. https://doi.org/10.1515/9781547401567-005

CLAI. (2015). *Publications – CLAI*. https://clai.ie/wp-content/uploads/2021/10/CLAI-Good-Practice-Discovery-Guide-v2_0.pdf

Cormack, G. V., & Grossman, M. R. (2014). Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 153–162. https://doi.org/10.1145/2600428.2609601

Current EDRM Model—EDRM. (2023). EDRM.Net. https://edrm.net/edrm-model/current/

IBM. (n.d.-a). *What Is Machine Learning (ML)? | IBM*. Retrieved May 7, 2024, from https://www.ibm.com/topics/machine-learning

IBM. (n.d.-b). *What Is Supervised Learning? | IBM*. Retrieved May 7, 2024, from https://www.ibm.com/topics/supervised-learning

IBM. (n.d.-c). *What Is Unsupervised Learning? | IBM*. Retrieved May 7, 2024, from https://www.ibm.com/topics/unsupervised-learning

Pace, N. M., & Zakaras, L. (2012). *Where the money goes: Understanding litigant expenditures for producing electronic discovery*. RAND.