Alexa, is this a historical record?

Santhilata Kuppili Venkata, Paul Young, Mark Bell and Alex Green
Digital Archiving
The National Archives, UK

Abstract

Digital transformation in government has brought an increase in the scale, variety and complexity of records, and greater levels of disorganised data. Current practices for selecting records for transfer to The National Archives (TNA), were developed to deal with paper records and are struggling to deal with this shift. This paper examines the background to the problem and outlines a project that TNA undertook to research the feasibility of using commercially available artificial intelligence tools to aid selection. The project **AI for Selection** evaluated a range of commercial solutions varying from off-the-shelf products to cloud-hosted machine learning platforms, as well as a benchmarking tool developed in-house. Suitability of tools depended on several factors, including requirements and skills of transferring bodies as well as the tools' usability and configurability. This paper also explores questions around trust and explainability of decisions made when using AI for sensitive tasks such as selection.

1 Introduction

It is the duty of Public Record Bodies under the Public Record Act (PRA) [1] to select records of enduring value for permanent preservation at The National Archives (TNA). Traditional processes designed to deal with paper records struggle to handle the volume, diversity and distributed nature of digital data. Motivated by this problem, TNA assessed the suitability and effectiveness of existing machine learning (ML) technologies for the selection of born-digital records (records created in an electronic form) for permanent preservation. This paper uses the experiences from the project to reflect on the challenges records managers (RMs) encounter and what role technology can play in meeting those challenges.

TNA has been investigating how technology can transform existing methods and aid government departments in undertaking appraisal and selection of their records. The TNA Digital Strategy [2] stated that 'we will develop new methods to help manage appraisal, selection and sensitivity review' adding 'we should investigate the use of machine learning'. Sir Alex Allan, in his Digital Records Review, states 'reviewers find it harder to scroll through data on a screen than to leaf through paper files' [3]. While this is referring to the process of sensitivity review rather than selection, it shows that in addition to challenges presented by the scale of digital material, the medium often renders traditional methods impractical. Machine learning (ML) has been applied to automated document classification, ranging from recent COVID-19 research papers [4], clinical records [5] to Brazilian legal documents [6]. Lee [7] notes that little attention has been focused on the use of computational methods for appraisal and selection in archives. Recent developments in ML have the potential to enable archives to deal with the increasing scale of digital records, ensuring the effective selection and preservation of historic public records for current and future researchers.

A study conducted by TNA in 2016 reviewed the use of eDiscovery tools to aid 'technology assisted review', including appraisal, selection, and sensitivity of born-digital material [8]. The current project **AI for Selection** deals with the selection issue exclusively. It aims at studying ML approaches and assessing the existing tools in the market to aid government departments in the selection process. It is perceived that any use of ML would aim to reduce the manual burden on RMs, still allowing them to retain the final decision on permanent preservation. The project studied a range of tools from off-the-shelf products to AI service platforms requiring code development.

The structure of the paper is as follows: the background and motivation to the problem that the project aimed to address is discussed in section 2. This section provides a detailed understanding of the issues around the transition from paper to digital records and how the scale and volume of digital records have impacted existing processes to select and transfer digital records. The outline of the project is detailed in section 3. The project used a sample of corporate records provided by TNA's Knowledge and Information Management (KIM) team to mimic data from government departments. The characteristics of the data are described in section 4. The help of multiple suppliers was sought to create or demonstrate workflows using existing technologies to handle the selection issue. An anonymised description of the solutions tested is explained in section 5. In order to evaluate the suppliers' products, TNA built a benchmarking tool with available open-source libraries. A complete step-by-step procedure of building such tools is given in section 6. An overall evaluation of the products is provided in section 7. Finally, a detailed discussion on the lessons learned and factors to be taken into consideration while developing intelligent solutions for sensitive tasks such as *Selection*, including the transparency of decisions made by AI, are discussed in section 8. This paper ends with a conclusion and suggestions for future work. There is a useful glossary of terms and acronyms in Appendix.

2 Background

The Better Information for Better Government (BI4BG) report [9] in 2017 described concerns which have arisen from the move from paper to digital record management.

> *"When information was predominantly held on paper, government was generally good at managing it. Files and filing were at the centre of how work got done: they were intrinsic to the flow of work, not an overhead on it. As a result, information could be organised and preserved and the life cycle from initial creation through to long-term preservation and presentation was robust."*

BI4BG noted that this was no longer the case with digital, referencing Sir Alex Allan's Review [3] which stated that the move from paper to digital documents and emails had 'undermined the rigour of information management across government'. While efforts by departments and BI4BG have worked to improve management of digital records there is a 'mass of digital data stored on shared drives that is poorly organised and indexed' [3]. Informal estimates by BI4BG put UK Government data at 16 billion emails and 3 billion documents amounting to 5PB of data, the equivalent of around 350 British Libraries, which is growing annually at around 9bn documents (120 British Libraries). In previous paper transfers between 2% and 5% of material is selected for transfer. For digital it could be argued that greater levels of duplication and ephemera will reduce this percentage further, but the whole is much larger. The result is a very large volume of poorly organised records that need to be appraised to determine if they contain anything of 'enduring value' that should be transferred to TNA.

BI4BG addressed the possibility of keeping everything, but they determined this was not feasible. The existence of personal data is one reason as retaining this information would 'contravene the principles in the Data Protection Act' [9]. It also stated that the ongoing management of material to ensure it remains 'usable and accessible' over time would come at a cost. While storage costs for digital material continue to fall, the savings are cancelled out by the growth in data volume. As well as financial cost, Pendergrass et al revealed the 'negative environmental impact' of 'ICT components, and therefore the digital preservation practices they enable' [10]. One of their recommendations was to embrace a paradigm shift when deciding material to select, critically examining content for 'enduring value' and considering the ongoing environmental cost as part of the process to help create 'sustainable digital preservation' [10]. Another argument against keeping everything is that if the task of selection is not undertaken at the point of deposit then the burden falls to the archive, or the user of the archive, to sift through the material instead. For paper records, Dunley, in his paper looking at the *Archive of the Edwardian Foreign Office records*, states that even if keeping everything was 'considered' desirable 'the scale of human information production means that archives can only ever contain a fraction of a once much greater whole'. Given the acceleration of information creation this is even more relevant in the digital age [11].

The current size of TNA's born-digital collections, excluding the UK Government Web Archive, in 2020 was around 31 TBs, and consists of a variety of formats. The PRA requires departments to undertake selection after 20 years[1] [1]. Current transfers come from an era that generated smaller volumes of digital data than today, as processes were still paper-based or staff followed print to paper policies. As a result, most departments are still within the limits of what can be achieved with manual selection processes, but volumes will increase rapidly as we move into, in archiving terms, *the new millennium*.

A survey of UK government departments preparedness for digital transfer in 2018 by Özdemir found that digital transfer was not yet a 'business as usual activity' for many departments [12]. This survey noted that some departments were unable to state when they would be due to begin their digital transfers. In many cases, this is because of the unreliability of file system metadata, which may have become corrupted, for example by system migrations, which makes it hard to date the earliest digital material. TNA has been exploring extraction of embedded date metadata within formats using tools such as Apache Tika, which often provides more reliable date information. It is also because even with digital records where departments are confident of their date until the appraisal is carried out the department will not know if any material will be selected for transfer to TNA.

Departments have been preparing for the digital selection challenge by developing processes for appraisal and selection of digital records and several have listed this as part of their Information Management Assessment Action Plans [13]. Often selection has been a manual process in classifying records for enduring value based on appraisal policies. Retention schedules are assigned to documents by records managers to define the life-cycle of a record (e.g retain for seven years). Documents selected for transfer to the archives are assigned a category for permanent preservation. These processes are proving impossible to scale to the volumes of digital material required to be reviewed. Without solutions to reduce the manual burden and deal with the scale of digital material, there is a risk that the transfer of material to TNA will stall. Hence there is an urgent need to find automated solutions.

[1]This period is transitioning from 30 years to 20 years meaning departments are currently transferring records from the late 1990's.

3 Project Outline

The aim of the **AI for Selection** project was to understand the existing market place for solutions to the born-digital selection challenge. Having already investigated rules-based eDiscovery tools, it was decided that this project should focus on ML. The intervening years since the eDiscovery project have seen an explosion in the capabilities and exploitation of ML. To the best of our knowledge commercial ML tools have not previously been evaluated for their suitability to aid selection of records, the time was right to look at tools using ML technology in the records management sphere.

A project was built to research the current state of the market for software products which could perform the selection task using ML. The project was designed to run in two phases, the first phase comprising exploratory desk research to understand the market and selecting at least 2-4 suitable products, and the second is to inviting suppliers to demonstrate their products' capabilities for a record classification task. A collection of documents from the organisation's Electronic Document and Records Management System (EDRMS) was curated, with associated contextual metadata, and provided to the suppliers to test their products against. The output of the project would be a report documenting the results of the testing and our assessment of the products for a government audience. The two phases were expected to run for 4 weeks and 12 weeks respectively.

In phase one, a consultant performed desk research to understand the market for ML-based technologies which could perform the record selection task, or partially perform the task, recognising a complete solution was unlikely to exist. The nature of desk research meant their assessments were based on the availability of detailed documentation on supplier websites. They were also asked to select a range of products demonstrating different approaches and capabilities. At the end of the investigation, five products were evaluated for the second phase.

In addition to selecting products, the consultant was tasked with defining assessment criteria to test the tools against. The assessment criteria included specific tests of the ML performance but also included functional aspects of the tools from an RMs perspective. TNA also decided to build their selection tool which served two purposes. First, it provided a learning opportunity for in-house data scientists to work with the documents and gain an appreciation of the challenges of record selection. Second, it would provide a base-level to compare the supplier tools against.

When briefing suppliers about the project we emphasised that the exercise was not a competition. The aim was not to achieve the highest ML accuracy but rather to demonstrate the functionality of their products to help government RMs understand the current state of the art and the trade-offs involved in choosing a product for the selection task. While each algorithm's performance was evaluated using standard accuracy metrics it is important to understand that these are not necessarily representative of a product's performance in general. The dataset was highly curated and belonged to one relatively small government department. Good performance against this dataset does not imply that equal performance would be achieved against the documents of a larger, more complex department. Although the project endeavoured to standardise the evaluation it will become clear later in this paper that comparison across products is difficult. For these reasons, we have elected to anonymise the products to avoid the reader drawing unintended conclusions about one product being 'better' than another. For the remainder of the paper, the products will be named according to single letters (A-E for the external suppliers and 'T' for TNA's solution).

The second phase of the project ran for a total of 12 weeks, with each supplier working for up to 8 of those weeks. Each paid supplier had the same budget which was to be spent

on people, licensing, and computation. The suppliers started and finished at different times but there were overlapping periods when they were all working. As a result the availability of internal staff had to be controlled and balanced across all suppliers. Interactions between internal stakeholders and each supplier were structured as follows: an initial meeting to discuss the project aims and objectives, and data; the data was then transferred to the supplier; fortnightly meetings to present early results and progress against the budget and final presentation of results. The output from each supplier was a final report detailing their experimentation, documentation of relevant features and workflow, and their responses to the requirements list.

4 Data

To evaluate supplier products, two sets of born-digital documents were curated. These came from two sources: the first set consists of records from Objective, TNA's EDRMS. This comprised files of various formats, but predominantly text-based, including emails, PDFs and Microsoft Office documents. These formats are representative of the majority of material produced by government departments. These files were organised in folders and sub-folders according to TNA department, and function or topic, as appropriate, following the KIM Team's guidance. The files have been individually labelled with retention schedules, however these labels were inherited from the label of their parent folder as set by the TNA KIM team. The sub-folder metadata included the retention schedule which specifies how long the records in the sub-folder will be kept for from the date the files were 'closed'[2]. As these files were from the corporate EDRMS, they were reviewed for sensitivities (personal or commercial) and a subset (118,677 files) suitable for sharing with external organisations was created. The files are referred to as the *'labelled data'* throughout this report. There were 20 types of retention schedules of which four (04, 06, 21, 33) were identified as records for permanent preservation. Their distribution is shown in Fig. 1. Each of the file types is shown with the volume of files 'not selected' for preservation in black and 'selected' type in grey. The second set of files were selected from a shared drive folder used by anyone at TNA working on content for the organisation's website. This comprised 50,697 files, mostly organised in folders but many within the root folder. This drive is considered a 'work in progress' area and is not actively managed or organised by any KIM Team guidance and no retention periods applied. They are referred to as *'unlabelled data'* in this article.

Two supervised learning tasks are possible with this data: classify documents by retention schedule; classify by preserve/delete. The labelled data were sourced from a well-managed file system so are not representative of the 'mass of digital data' that Sir Alex Allen referred to, but for testing ML they sufficed as a training set. Since they were well organised, a classification algorithm using solely the file and folder metadata would be

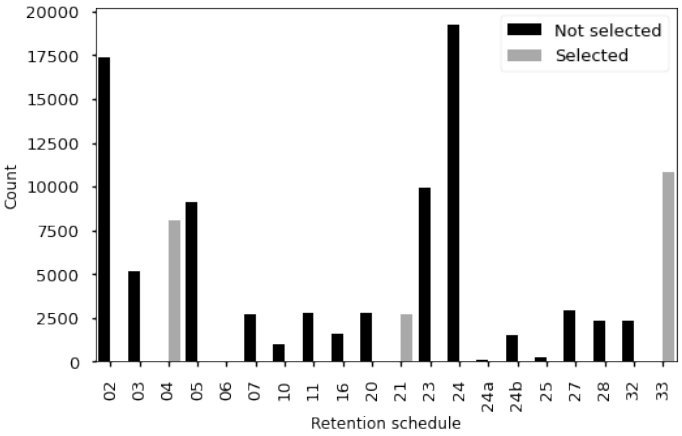

Figure 1: File distribution according to retention schedule

[2]Files that are no longer actively in use by the business

expected to achieve very high
levels of accuracy. This hypothesis was tested by *product T*. However, it would not be a useful test as similar performance could be achieved without ML and the problem under investigation was the classification of unorganised records. By not passing folder and department metadata to the ML, the task is rendered indistinguishable from labelling unorganised files. The suppliers were therefore asked to only use the content of documents to perform classification. The labelled data was the primary dataset for the project, while the unlabelled set was given lower priority.

5 Products

This section details the products and their features.

Product A

A cloud-based Software-as-a-Service (SaaS) platform which is designed to enable RMs to use ML without the input of a data scientist. Training a model is achieved by sampling the labelled dataset and manually classifying the sampled documents through the GUI. Following training, unlabelled documents are imported. De-duplication and named entity extraction are automatically applied as documents are loaded, with the extracted entities being used as features for the ML. The newly loaded documents are given suggested labels by the ML classifier but they are not final until they have been approved or corrected by the records manager (RM). The model can be iteratively re-trained as the user works through this process.

Product B

A cloud-based file analytics platform. Although the platform is designed for ease of use by an RM, it also provides functionality for data scientists to engage with the model building process. The product has a three stage training process involving unsupervised clustering, entity and NLP-based feature extraction, and a semi-supervised refinement using labelled data. It was not clear from their report what model was used or the amount of parameter tuning undertaken, but their main approach to influencing results was manipulating training samples. This makes sense for a system primarily designed for an RM over a data scientist.

Product C

A content-services-platform centred around records management with the ML capability provided by an automated-cloud-ML service. Their business model was for the ML training to be a managed service performed by their own data scientists, rather than the user. The interface was therefore focused around organising and searching documents rather than the ML process.

Product D

A cloud hosted ML as a Service (MLaaS) platform. A model is trained through labelled examples but model selection and parameter tuning are opaque and not configurable. The platform itself includes open ML algorithms but the provider were specifically asked to test this particular service. The documents and their classifications are loaded through an API which requires some programming skill to implement. A full pipeline can be built programmatically using the platform.

Product E

A cloud hosted MLaaS platform with a number of algorithms available and the ability to build an ML pipeline. A prototype GUI was also created to demonstrate how an RM could execute workflows on the platform.

Product T

The benchmarking tool is written by TNA team, described in section 6 as a worked example of the process of applying ML to document selection.

6 ML Pipeline for Benchmarking

This section details the development of the benchmark tool, *product T* and discusses the general methodology, data pre-processing, and ML training and evaluation. It was created using the Python programming language and free open source ML libraries, with a command line interface (CLI) rather than a GUI. Although the solution requires technical skill to use it, the problem and resulting ML pipeline were modelled from a RMs perspective.

6.1 Methodology

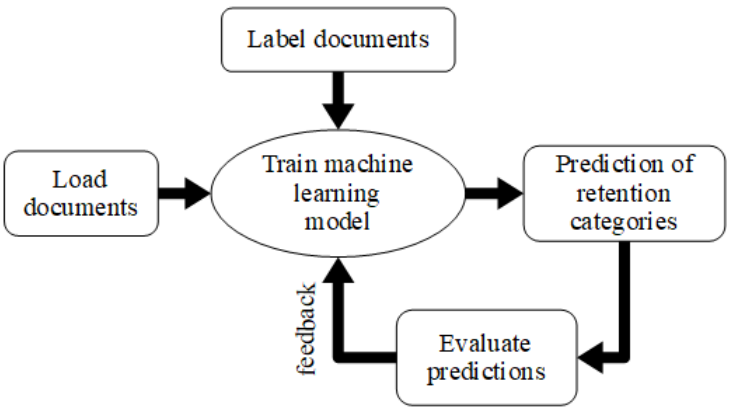

Figure 2: Methodology of machine learning pipeline

The methodology is divided into five activities as shown in the activity diagram in Fig. 2.

Load documents: Since the benchmarking tool was developed within the secured environment of TNA it loads files from a local drive using the CLI. In a production system, the files could be loaded from a local server, external drive, EDRMS, or cloud storage and all of the supplier products offered this functionality.

Label documents: Labelling records with the correct retention category is how training data is provided to the ML model. The labels and supplementary metadata were loaded directly from the spreadsheet.

Train ML model: Supervised ML algorithms work by learning from labelled examples of data, where each example is made of a set of features which have been extracted from source data. An example of a feature when working with text would be a dictionary word. The algorithm then learns to weigh each feature in such a way that it minimises the number of new examples it classifies incorrectly. Feature engineering and the algorithms used by *product T* are explained in later subsections. Two of the products used algorithms opaquely selected by the automl process of the cloud provider. A product used a rule-based process in conjunction with ML, while another experimented with various open source ML algorithms. The 'configurability' of a product is mainly decided by this process in the workflow.

Prediction of retention categories: This activity predicts labels (retention categories) for previously unseen documents which were excluded from the training process. The *product T* used a two-step process firstly predicting retention schedules and then mapping them to a binary ('Selected' or 'Not selected') value for permanent preservation decision. The supplier products either predicted retention schedules, or used a binary classifier (not derived from retention schedule).

Evaluate predictions: The RM can inspect prediction results and fine tune the predictions through configuration settings. For example, to select a different model, or to change the training data. However, the efficiency of the fine tuning depends on the skills of the user. The evaluation requires data science skills to interpret accuracy results and change configuration, and a clear evaluation scheme, and this was the case for any of the tools which had configurable ML.

6.2 Training and Test datasets

In order to test the effectiveness of an ML model it is trained with one set of data and then tested with a second. Since only one dataset was provided, two lists of record identifiers were sent to suppliers to standardise the training/test split. In a competitive scenario the labels for the test set would have been held back from the suppliers. Suppliers were also asked to use 10-fold cross validation which generates 10 training and validation sets and averages the results. Cross validation shows whether algorithms generalize well when presented with new data. How to split the data with each iteration is important too. Since the classifications were highly imbalanced a weighted approach to splitting data was used to make sure every class was represented in the training data. The danger of not doing this is that if a class does not appear in the training data it will be unknown to the algorithm.

6.3 Data Analysis

There was no missing data in the metadata. However, the contents of the documents needed to be cleaned prior to training the model. The labelled data was a mixture of both text documents and media files with 143 distinct file types in total. The rules used by KIM to set the document retention schedules are generally applicable to text documents only. The dataset was therefore restricted to 95,402 text documents with extensions *.doc, .docx, .rtf, .pdf, .txt, .msg, .xls, .xlsx*. The Fig. 3 shows volume of documents by file type and label (Selected/Not Selected) in the dataset. It can be seen that the dataset has a large number of .msg files (emails) with a large portion selected for permanent retention. It indicates that emails stored in the EDRMS are highly likely to be picked up for permanent preservation, possibly as an evidence of decisions being made. Other important file types include .doc (.DOC and .docx) and .pdf and .rtf. Further

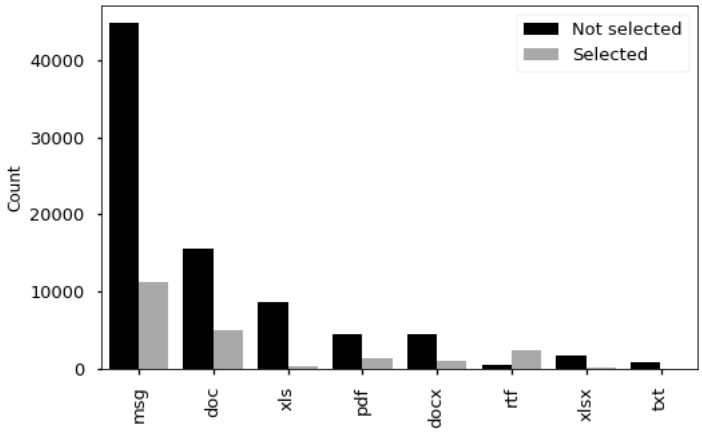

Figure 3: File distribution according to file types

analysis pointed out a correlation between selection and departments and selection and top folder. However, we are not presenting those statistics here due to data anonymisation.

6.4 Problem Modelling

The data exploration indicated that a classification model could be built on the metadata alone since there was a strong correlation between some of the features and the preservation decision. However, as explained in section 4, it is unlikely that this level of metadata would be available in a typical collection from the late 1990s. So two possible approaches were tested with the benchmark tool:

1. Predict the retention schedule from document metadata,
2. Predict the retention schedule from documents contents.

This allowed comparison between ML in an organised environment (using metadata) and scenario where we do not have metadata available.

Supervised Vs Unsupervised approaches for modelling

The dataset received for the task is well-defined labelled data. Hence the methodology in subsection 6.1 presumed the task to be a multi-class classification (supervised task) of documents into various retention categories which are then mapped to binary values (selected/not selected). Document classification also could have been performed as an unsupervised learning task by grouping documents into clusters of similar content. One reason for not considering the unsupervised methodology for this problem is due to the class imbalance present in the data. While some classes are heavily represented, other crucial classes were sparse. The process of document clustering based on the contents such as, t-SNE and K-Means algorithm combination[3]) missed those under-representative classes altogether. All suppliers' products performed supervised learning only. This decision partially due to the well labelled dataset. Also we did not see evidence that any of the tools had unsupervised learning capabilities.

[3]These two algorithms are commonly used in combination on text data for clustering

Another issue to keep in mind while using classification models for selection problem is, the *false negatives*. Especially for the selection of sensitive documents, false negatives prove to be more dangerous than false positives. A false positive is a document that is classified as a 'selected' category while actually it belonged to 'Not selected' category. Whereas the false negative is an essential document classified as 'not selected' category, which should be in the 'selected' category.

6.5 Classification using Metadata

In addition to the metadata supplied by spreadsheet, Apache Tika extracted extra metadata. The metadata features selected for model training were related to the file type, repository from where the document was sampled, author, file size, retention schedule for preservation, version number, time last modified, and top parent folder. On further examination, author and time last modified fields were omitted as they were overwritten and corrupted while transferring data from its source to the experimentation site. To develop a model using only metadata, the benchmarking tool explored *Naive Bayes* [14] and *Decision tree* classification [15] algorithms. The algorithms were chosen for their simplicity and their explainability with respect to relating results to feature characteristics, and also they were expected to set a useful baseline for product evaluation.

Naive Bayes is a simple and efficient prediction algorithm which performs well in multi-class prediction. It makes an assumption of independence between features, i.e. a change in feature 'X' has no effect on the value of feature 'Y' and when this holds, a Naive Bayes classifier performs well in comparison with other models. The benchmarking tool obtained an accuracy of 63.2%. No further feature engineering was done but these results suggested that strong dependence between features which has negatively affected the results.

A **Decision Tree** classification model closely follows a human decision making process by splitting the data one variable at a time. Implementations of the algorithm are often able to work with both numerical and categorical data with very little pre-processing. For complex tasks it is often outperformed by other more sophisticated approaches but it performed well for this task. It is the most interpretable algorithm and ideal for situations where a decision must be explained clearly and a person could use it to exactly reproduce the results from the algorithm. This model obtained an accuracy of 93.4%.

The decision tree model has classified unlabelled data in to three retention categories ('16', '23' and '33'). Of these majority of documents are classified as '23' which is partially a satisfied result as these documents belong to informal projects in general. On manual inspection, TNA project team found that all 32 documents that are classified as '33' (category for permanent preservation) achieved 100% correctness as they should be in selected category.

6.6 Classification using Contents of the Document

In this section, we discuss the model developed by using the contents of the documents as features. In ML, document classification is done with the help of natural language processing (NLP) but for the selection of documents, the content is more important than linguistic features [16].

6.6.1 Document data pre-processing and feature extraction

ML algorithms generally require numeric inputs and the textual content needs to be converted to features in a numeric form. Depending on the application, the features may be individ-

ual words, multi-word phrases (n-grams), linguistic labels (verbs, nouns etc.) or even topic models[4]. While designing NLP applications for document classification, feature extraction therefore becomes a significant part of the development. Text documents follow complicated rules defined by the language while conveying implicit meaning, but the ML models we used are based on mathematical and statistical principles which are not nuanced enough to capture the meaning and understand the text. For this reason, ML can still be a blunt tool when compared to a human expert. Expert led feature engineering combined with simpler ML algorithms can still perform well.

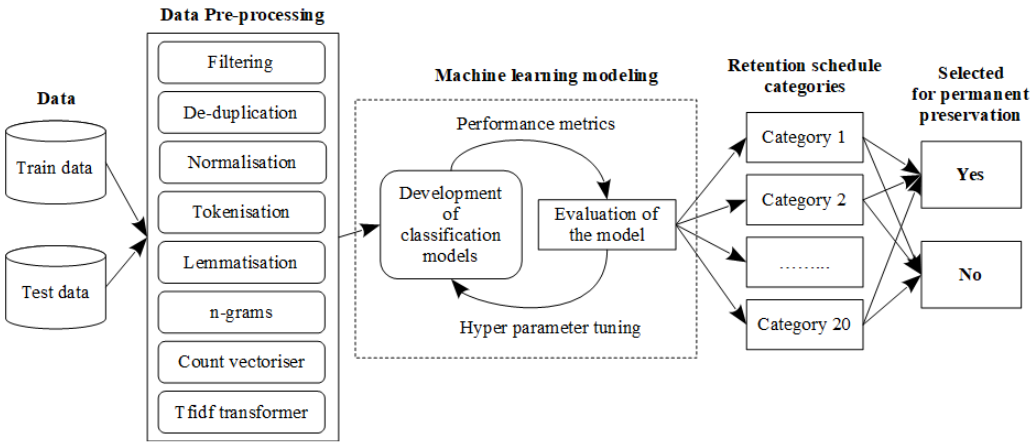

Figure 4: Pipeline to select documents using NLP on data contents

The Fig. 4 shows the complete ML pipeline used for developing the *product T*. Train and test datasets are loaded in to the data pre-processing phase. Both datasets go through the same data processing steps before being used by the rest of the pipeline. The feature extraction and pre-processing steps are listed and described in further detail in appendix A.1. Next, the processed features are given as input to ML models. Training a model and testing on the test data is an iterative process until the results are obtained. The *product T* used a two step classification; First classify into various categories and then select the preserve/delete based on the type of retention category. Three classification models: Naive Bayes, Random Forest and Logistic Regression were tested using the above pipeline[5].

Table 1: Performance metrics

	Naive Bayes				Logistic Regression			
Class	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Not selected	0.83	1.00	0.90	14471	0.97	0.98	0.97	14471
Selected	0.98	0.30	0.45	4296	0.93	0.90	0.91	4296
accuracy			0.84	18767			0.96	18767
avg	0.86	0.84	0.80	18767	0.96	0.96	0.96	18767

[4]A topic is represented as a distribution over words, and each document is then represented as a distribution over topics

[5]Random Forest performed similarly but not as well as logistic regression and is omitted for brevity

6.6.2 Model 1 - Naive Bayes classifier

For this task, the Naive Bayes classifier performed comparatively well in terms of overall accuracy (84%) but suffered with poor recall (30%). This means that of all the documents in the dataset which should be preserved, it correctly classified 30% of them only. The precision for this class was excellent, being correct 98% of the time out of all documents marked for preservation. Results indicate that there are more false negatives compared to false positives as shown in Table 1.

6.6.3 Model 2 - Logistic regression classifier

The **Logistic regression** is a statistical model which can perform multi-class prediction. While handling millions of features (words and phrases), a logistic regression model [17] allows us to fine tune the parameters to suit the needs due to the probability theory behind the algorithm. Unlike Naive Bayes, it can account for dependencies between variables. Multi-class performance metrics are not provided in this paper due to space constraints. The logistic regression model outperformed Naive Bayes (96%) accuracy) and was far more balanced in terms of precision (93%) and recall (90%) for the 'Selected' category[6]. Overall evaluation is detailed in section 7.

7 Evaluation of Products

The TNA project team were asked to rank the products according to their usability and configurability, from the perspective of an RM and ML functionality of the tool. High usability would suggest a tool which required a low skill level to prepare training data and train a model, while a highly configurable tool would give the user a lot of influence over feature engineering, algorithm selection and fine-tuning of the algorithm. Ties were allowed in the rankings, and two (out of six) people rated *product C* as N/A for configurability. This rating reflected the fact that although the user had no direct control over the training process, the model building was undertaken by outsourced data scientists and therefore a high level of configuration would occur but it was ambiguous how to rate it. The rankings were averaged and the results are shown in Fig. 5. No product scored highly for both measurements.

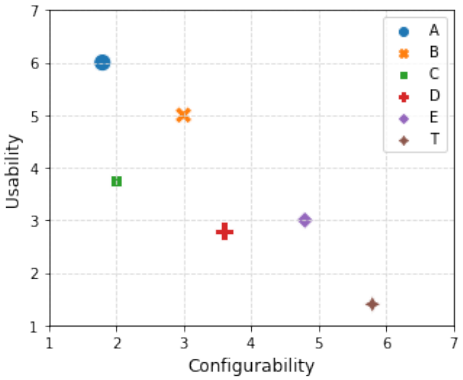

Figure 5: Project team rankings of products

Fig. 6 shows boxplots of precision, recall and F1-score for all tools. Precision is a measure of how likely a classifier is to be correct in its classification (or how much a user should trust its answers). Recall measures the percentage of positively labelled records ('preserve' in this case) which have been correctly classified. Poor recall (like the 30% result for Naive Bayes earlier) would mean many important documents would not be selected for preservation, while poor precision would mean many non-important documents would be needlessly archived. The F1-score[7] is a balanced combination of precision and recall. Each plot is overlaid with points

[6]Github link for the source code of *product T* is available on request

[7]Mathematically F1-score is the harmonic mean of precision and recall

representing the actual values summarised by the boxplots and coloured according to their configurability ranking (in Fig. 6). The recall scores are very wide ranging compared to the precision due to the minimum and maximum values being 31% apart (versus 19%), although the interquartile range for recall was narrower than that for precision. The low recall of *product A* may be due to the low volume of training data it used. The three best performing products (by F1-score) all achieved over 90% precision but only one passed that mark for recall. While this level of recall may be a concern we lack any figures for human performance to compare with, so it is difficult to assess what is 'good' performance. The coloured points shows that the tools which were highly configurable (including *product C*) performed much better than those which low on that scale. Surprisingly, the in-house benchmarking tool with a simple logistic regression model outperformed all of the others. The implications of these results will be discussed further in the section 8.

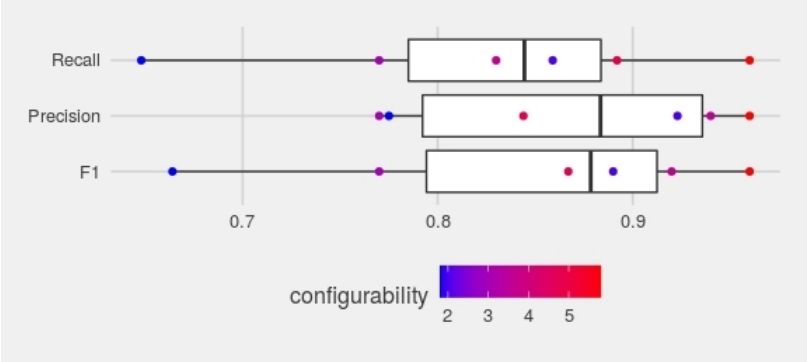

Figure 6: Performance metrics in relation to configurability ranking

As discussed in 6.2, evaluation was performed against a standard test set. However, several challenges arose in comparing the results of products. Not every supplier used all of the data supplied for training. *Supplier A* was limited by the number of records which could practically be manually labelled through their interface, while *suppliers D & E* had to control their cloud usage to remain within the budget. This difference may explain why *product A*, using only 6% of the data, significantly underperformed on the F1 measure (¡0.7) compared to *product T* (0.96) which used 96% of the data. If we consider the realistic situation of a disorganised collection of over 100,000 documents which needs to be labelled by RMs then it is more reasonable to expect them to manually classify a small percentage. If labelling tens of thousands of documents based on their content alone was feasible then perhaps there would be no need for ML in the first place. Research into sensitivity review by MacDonald [18] posed the problem as one of balancing automation and limited human resource. This is a sensible direction for future product development incorporating active learning and semi-supervised ML technologies so that the RM and machine work in tandem, rather than current disjointed approaches involving bulk labelling and training. Suppliers using cloud ML platforms found that their budget was challenged when they tried to process large volumes of data, and so restricted the number of records they used for training. One discovered that a tool for extracting text from images was proving expensive so they streamlined their pipeline to prevent overuse of that product. Another found that there was a small, but significant, the number of documents which were very large in comparison to the average file size across the corpus. By removing large files, they were able to reduce costs considerably but it raised questions of how large files should be processed if the cost is an issue. One final reason for not using all the data was that some suppliers were either not able to process, or just concentrated on, certain document types (e.g.

word-processed documents and emails).

8 Discussion

The products have been classified according to usability and configurability, each of which raise points for discussion. Following are some of the valuable lessons learned during the project.

Enduring Value

A problem with moving to ML solutions is adapting processes for accurately assigning enduring value which has often been a subjective and shifting classification. In terms of digital strategy practice at TNA, much would still be recognisable to Jenkinson [2, 19]. The government departments, as the creators of the records, select material based on two broad criteria: 'the documentation of what government did, why and how' and 'the value of the records for future historical research' [20]. Departments create and publish appraisal policies which outline how decisions are undertaken and what material will be selected [21].

Significance

Ideas of historical and archival significance will likely evolve [22]. As an example of changing opinions on value, the Grigg committee (1953) said that the government should not preserve records of genealogical or biographical value, partly due to worries about storing the volume of the material concerned. This led to sampling procedures for some collections which meant only a portion of records were transferred, often containing large degrees of unintended bias. The ever-growing interest in family history, along with possibilities of digitisation led to a rethink, especially in the case of the National Health Service Central Register. To avoid increasing any bias in previous decisions, any system or dataset would also need to be able to evolve. There need to be methods for making selection decisions to identify incorrect decisions and improve future models. One method to address past bias could be to assign more weight to recent decisions, functionality which was built into one of the products. For digital records, there is also a question of format. New technologies mean that the concept of what is a 'digital record' is always expanding. As increasingly more algorithms are used to make decisions in government, they too become a potential record to be selected for archiving. Many approaches detailed in this paper were focused on text-based formats that are held in a single file. Certain formats are not appropriate for techniques used in this paper. This includes documents which may be covered by multiple files, such as 3D modelling formats and HTML web-pages, or non-text based formats. Volumes of this material do not currently require machine learning approaches to determine selection.

Data Quality

It is clear from what we have seen that the quality of the training data is critical to good results. While feature engineering and model tuning are an important aspect of optimising results, *supplier B* reported that selection of training data had the biggest influence on accuracy. Even those products promising a user experience geared towards a RM still embedded data scientists in their project teams suggesting we are still a long way from not requiring skilled data professionals in the process. This point is emphasised by the fact that every supplier began with a phase of data analysis and data visualisation, functionality which was not built into the tools themselves but is recognised as an essential part of the ML pipeline.

Training models

Traditional software products are designed from a set of requirements gained through user engagement. The expert systems in the 1990s were created by eliciting logical rules from experts. But ML is different. Training a supervised learning algorithm for selection means trying to imprint the context of an appraisal policy via the training data. Addressing concerns in data preparation is not a concern unique to archives, Jo et al. suggested that ML pipelines could be enhanced by improving data collection by using some of the methods undertaken by archives for appraisal namely, recording the 'process of data collection' and relying on multi-layered, and multi-person systems rather than 'a single ML engineer' when compiling a dataset [23]. Our view is that records management teams should be equipped with tools for data mining and analysis, and the requisite skills to use them. They must go through the process of curating training sets of documents which effectively represent the rules and processes they follow when selecting documents, before jumping into ML. To the uninitiated the ML may sound difficult and complex but it is not, as the data is the hard part rather than ML. At TNA we have begun a series of internal workshops to explain and demystify the ML process for non-technical staff, and we would recommend similar training for government departments [24].

Disorganised data

In real-life applications data often comes disorganised, like the unlabelled data we received for the project (refer to section 4). This unlabelled data suffered from two issues. First, document folders are nothing but the containers for a mixture of file types, and non-standardised file structures belonging to various departments dumped together. This scenario is relatable to common shared drive data in any organisation. It needs a thorough cleaning of data by data specialists and domain experts working together. Domain experts are needed to use their knowledge to provide some order and to identify high-level document features which can aid selection. Data specialists need to perform *feature engineering* to extract and generate features from the documents at scale to be fed into a ML model.

The second issue is with data representation. While performing classification with machines, it is necessary to have data points with good representation in each category. Data are said to suffer the *class imbalance problem* when the class distributions are highly imbalanced [25] in large data volumes. In our document corpus, the class imbalance is shown in the fig 1. While there is a majority representation for categories: 02, 04, 05, 23, 24 and 33, there are hardly any files for classes: 06, 10, 24a, 25. Such huge class imbalance induces *bias* in the decision making. A serious problem in the given data is retention category 06 which should be categorised as 'Selected'. But with a negligible representation in the training data, the ML model might fail to recognise that category at all! As a result, there is a chance that all documents in that category will be misclassified as 'Not selected'. Theoretically, there are few ways to handle the imbalance problem [26, 27, 28].

- *Over/under sampling of the minority or majority classes:* Sampling techniques are statistical methods, work well for numeric data but we found hardly any difference with document data.
- *Data augmentation:* Data augmentation is a useful technique to create synthetic data for image analysis but cannot be applied to documents.
- *Intelligent feature selection:* By selecting only those features that are common to all other categories. This approach might give good results With fewer features. We need to experiment with this approach in future.

- *Algorithmic approach:* We have experimented with the **bagging** algorithm along with Naive Bayes and logistic regression models for *product T*. Being an ensemble model, the random forest model (a type of bagging model) [29] performed as well as Logistic regression.

Automated Machine Learning (automl)

There is a trend towards automl which aims to automate the process of selecting, and tuning the hyperparameters of ML models. This is a useful innovation in two respects. For the data scientist, this part of the process is often one of educated trial and error so is suited to automation, and there is evidence that automl can outperform humans for some classification tasks [30, 31]. For the organisation there is a potential cost saving since the 'democratisation of machine learning' offered by automl means pipelines can be built by existing developers, rather than more expensive and harder to recruit data scientists. However, hyperparameter tuning is only a small part of the data scientist's role. The CRISP-DM [32] process model is a commonly used methodology for designing a data mining project life-cycle, and there have been attempts to create a ML equivalent [33, 34]. What these methodologies have in common is that modelling requirements and understanding data are a large part of the process and are not ready to be automated. Additionally, the evaluation phase and on-going monitoring require data science skills, otherwise the automl is effectively marking its own homework. Highlighting the complexity of putting a ML pipeline into production, Breck et al. have proposed a 28 step ML readiness rubric [35]. Ongoing monitoring is a part of the rubric that includes tests for model 'staleness' which can occur when the distribution of incoming data changes over time. One of the products we tested included the facility to weight training data according to its age.

Cost models

The range of tools used demonstrated four different cost models (licensed, outsourced ML, MLaaS platform, in-house development) which will be a big factor influencing the selection of a solution. Three of the products were licensed records management products which incorporated ML technology to aid selection. The cost effective way to use them would be for an organisation to migrate their records to the product, but the choice of a document management system depends on more than its record selection functionality. Of these three, one of them offered the ML as a service. While this will be attractive to an organisation which does not have access to in-house data scientists, it will be difficult to scale this offering if it becomes popular. The two MLaaS cloud platforms offer a component based approach to constructing a record selection pipeline. Record transfers to TNA tend to be performed in annual batches (although this may change as digital transfer becomes the norm) and cloud offers benefits in terms of being able to scale up the application when large transfers are initiated. However, the experience of the suppliers highlighted concerns with the costs of the ML aspects of the process. We would expect these to reduce over time as the products mature but departments with millions of documents may find the costs prohibitive at the moment. The fourth option would be building an in-house solution which could still run on cloud servers, taking advantage of the scalability, but use open source libraries rather than MLaaS. However, the advantage of the MLaaS offerings is that they reduce the development effort considerably as they involve stitching together ready made components, with some customisation, rather than developing from scratch. The automated ML approach reduces the development effort further by removing the need to evaluate different ML models, but it is more expensive to train models. None of the 'build your own' ML products had a GUI and so the other products have a distinct advantage in

terms of usability for a RM. Without significant investment in a usable interface records management teams will require embedded technical staff to run their record selection processes, which would be a big change for most departments.

Risk

Another aspect of incorporating ML is the treatment of risk. The products were evaluated according to precision (the level of trust one should have in a positive classification), recall (the percentage of true positives correctly identified), and F1-score (a combination of precision and recall). While ML systems are often evaluated using the F1-score, in the case of record selection we may want a system which is biased towards either precision or recall. Consider a system which has very high recall but which archives 20% of a department's documents instead of the usual 5%. This would place a heavy burden on the archive in terms of long term storage costs, the teams responsible for cataloguing the archived documents, and on the user of the archive who has more ephemera to sift through to find interesting material. The mitigation therefore would be to put more resource into manually selecting the excess material. Alternatively a system with high precision but lower recall would require less manual work, but the danger of not archiving important documents is increased. Of course, high precision and recall would be ideal but until that comes along the departments need to understand their own risk appetite and resource availability. TNA has experience in eliciting quantification's of risk [36]. Another risk in using ML systems is that they generally involve data migration, moving the data into areas where the tools can analyse them. This creates potential for corruption of files or metadata, as well as introducing sensitivity concerns when uploading documents to third party cloud platforms. A Document Management System which incorporates the ML element avoids this issue.

The context of decisions made

Understanding the decisions made around selection of records is important not just to RMs who make selection decisions but also to the long-term understanding by future researchers of the collections. Dunley states, that to understand the archive fully you must understand the processes through which 'the selection and preservation of "valuable" information occurs' [11]. Understanding existing decisions relies on analysis of appraisal policies and procedures. The addition of ML approaches adds additional complexity to this understanding. Manoff has listed concerns around 'impenetrability of machine processes and algorithms' in her study assessing the potential for technology to create areas of 'archival silence' [37]. While algorithms do not have to be impenetrable some are, especially proprietary ones using automated ML approaches. It could also be argued that human based selection which involves interpretation of policies is also impenetrable, we are not always able to request the reason behind a record's selection.

Explainability

As ML makes more decisions affecting our lives, the issues of transparency and explainable AI (XAI) have come to the fore in the fields of Human Computer Interaction and AI. Following a workshop on Human-Centred Explainable AI, Bunn offers some reflections on XAI from a records management perspective [38]. Algorithmic decisions can be explained at the level of the model (e.g. how does it function? What assumptions does it make?) or at the individual record level (why was this record classified as A rather than B?). The first type of explanation can be achieved through transparency by using open source code and through documentation.

Of the six products only two achieved this level of transparency and none of the products went to the level of explaining individual predictions. The lack of explainability reflects the fact that the ML functionality in these products is still new and experimental. It is debatable whether knowing the specific algorithm is important to justifying the automated selection process. The choice of algorithm is generally a technical decision based on its suitability to the task and data. However, the choice may also be influenced by the need for explainability and this needs to be weighed against accuracy. For example, a decision tree is highly transparent and can be understood by most people after some initial explanation, while a deep neural network would be impenetrable to all but experts in the field. The neural network will, however, generally outperform the decision tree. Explaining individual decisions is more important and there are a number of techniques available such as LIME [39] and Shapley Values [40]. Both of these methods provide a way to understand which features have most influence on the model's classification of a record. All of the suppliers provided detailed statistics for both overall model (select/not select) accuracy and individual class (retention schedule) level. These statistics were presented as both tables and confusion matrices. One supplier also provided the Brier score[8] which is a measure of how accurate the classification probabilities were. If a classifier returns a confidence of 0.8 then we would expect that is correct 80% of the time, but if it were only correct 50% of the time then we would be less trustful of the confidence score. This is important in a situation where there is limited resource to quality assure results. If the probabilities are accurate then the record manager can save time by trusting the 99% confident results and focus on the lower probability ones. If they are not accurate then it is less clear which results to prioritise, and trust in the system will be reduced.

Conclusion

The selection of born-digital records for permanent preservation is clearly a problem for government departments and they will need technological solutions to help sift and process the large volumes of data. ML has the potential to be an important part of the solution and with sufficient labelled data we would expect it to perform well. The challenge is that labelling data is resource intensive and the nature of government records means it can not be crowd sourced. It is also clear that ML still very much requires data scientist skills and records management teams either need to incorporate those skills, or they need to engage with the existing data science community in government. That community are already in high demand and so it will come down to priorities, but compliance with the PRA should have some influence. As well as technical skills there also needs to be an educational programme for non-technical staff to understand the concepts of ML and their role in creating the data to train the system. Our view is that a suite of tools drawing on data mining, human computer interaction, and AI technologies, that can interface with repositories of records, rather than a single solution will be needed. Record selection is about enacting a policy and requires knowledge of the collections and events in the outside world. This requires experts leveraging technology to make their job possible, rather than relying on it to perform the selection task for them. These tools are labelling at the document level, not at the folder level, and do not take into account the context of documents as a collection within a series or folder. Maintaining human control over the process can allay fears of machines making decisions. Transparency is vital and more work is needed to explain why algorithms are making decisions and align them with policies. The processes and rationale behind training data curation should be published to help identify potential biases. The good news is that all of the products we saw were in early stages of development which

[8] https://www.jclinepi.com/article/S0895-4356(09)00363-1/pdf

means this is a great time to engage with suppliers and influence their future development so that they work for RMs and archivists.

Future Work

In this project we have worked with labelled data which steered all of the solutions towards supervised learning methods to solve document classification. However, the problem that needs to be solved is when the data is disorganised and unlabelled. This is partially achieved through good interfaces for exploring the records, and some of the suppliers have made good progress in that area. However, the volume of data means there are limits to what filtering within a GUI can achieve. Recent advancements in NLP such as deep learning using transformer models [41] and unsupervised document clustering with attention models, both of which use word context, have achieved impressive results over the last couple of years [42]. Document selection is still a nuanced, expert activity, and so methods for combining ML with human defined rules should also be explored. At the moment none of the products support combining rules with classification ML techniques. Further research into classifying documents as a collection, and combining ML with context and human knowledge is needed. The TNA project team's ratings of usability were based solely on supplier presentations rather than hands on experience of the products. A more extensive project would assign more subject matter expert resource to experimenting with the products themselves and this will be an important aspect of any future work. By working with these products records management staff can understand more clearly how the tools would integrate into their workflow and gain experience of curating training data, building and evaluating models, and working with AI driven assistance. It would also be an opportunity to explore the trade-offs between usability and configurability. With the products that were evaluated there was a clear distinction between those which provided an intuitive GUI for working with documents, and those which were aimed at the data scientist and controlling the learning process. None of the products offered both. It is difficult to achieve both without either automating the work of the data scientist or embedding ML skills within the records management teams.

Acknowledgements

To our entire project team: Michael Appleby, Jérémie Charlet, Nicola Welch, Bálint Csöllei, Claire Driver, Susannah Baccardax, John Sheridan, the Procurement department (TNA), and the suppliers without whom this project would not have been possible.

References

[1] Public records Act 1958. Public records act 1958.

[2] The National Archives. *The National Archives Digital Strategy 2017-2019*, March 2017. `https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf`.

[3] Sir Alex Allan. Review of government digital records, 2015.

[4] Hamed Hassanzadeh, Anthony Nguyen, Sarvnaz Karimi, and Kevin Chu. Transferability of artificial neural networks for clinical document classification across hospitals: A case

study on abnormality detection from radiology reports. *Journal of Biomedical Informatics*, 85:68 – 79, 2018.

[5] Bernal Jiménez Gutiérrez, Juncheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. Document classification for covid-19 literature, 2020.

[6] Nilton Silva, Fabricio Braz, and Teofilo de Campos. Document type classification for brazil's supreme court using a convolutional neural network. pages 7–11, 10 2018.

[7] Christopher A Lee. Computer-assisted appraisal and selection of archival materials. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2721–2724. IEEE, 2018.

[8] The National Archives. The application of technology assisted review to born-digital records transfer, inquiries and beyond, 2016.

[9] BI4BG. Better information for better government, 2017.

[10] Keith L. Pendergrass, Walker Sampson, Tim Walsh, and Laura Alagna. Toward Environmentally Sustainable Digital Preservation. *The American Archivist*, 82(1):165–206, 03 2019.

[11] Richard Dunley. The archive of the edwardian foreign office: The archaeology of a collection and its use. *Diplomacy & Statecraft*, 31(3):429–449, 2020.

[12] Lale Özdemir Şahin. The inevitability of digital transfer how prepared are uk public bodies for the transfer of born-digital records to the archives? *Records Management Journal*, pages 224–239, 2019.

[13] The National Archives. Ima reports and resources.

[14] Geoffrey I. Webb. *Naïve Bayes*, pages 713–714. Springer US, Boston, MA, 2010.

[15] Johannes Fürnkranz. *Decision Tree*, pages 263–267. Springer US, Boston, MA, 2010.

[16] Dunja Mladeni, Janez Brank, and Marko Grobelnik. *Document Classification*, pages 289–293. Springer US, Boston, MA, 2010.

[17] Claude Sammut and Geoffrey I. Webb, editors. *Logistic Regression*, pages 631–631. Springer US, Boston, MA, 2010.

[18] Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins. Towards a classifier for digital sensitivity review. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, pages 500–506, Cham, 2014. Springer International Publishing.

[19] H. Jenkinson. *A Manual of Archive Administration*. Carnegie Endowment for International Peace: Division of Economics and History. P. Lund, Humphries & Company, Limited, 1937.

[20] The National Archives. Best practice guide to appraising and selecting records for the national archives, 2013.

[21] The National Archives. Operational selection policies by subject.

[22] Valerie Johnson, Sonia Ranade, and David Thomas. Size matters: The implications of volume for the digital archive of tomorrow – a case study from the uk national archives. *Records Management Journal*, 24(3):224–237, 2014.

[23] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery.

[24] Mark Bell, Leontien Talboom, and James Lappin. Machine learning club. IRMS, 6 2020.

[25] Charles X. Ling and Victor S. Sheng. *Class Imbalance Problem*, pages 171–171. Springer US, Boston, MA, 2010.

[26] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano. A comparative study of data sampling and cost sensitive learning. In *2008 IEEE International Conference on Data Mining Workshops*, pages 46–52, 2008.

[27] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

[28] M. Wasikowski and X. Chen. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1388–1400, 2010.

[29] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[30] Marc-André Zöller and Marco F Huber. Benchmark and survey of automated machine learning frameworks. *arXiv preprint arXiv:1904.12054*, 2019.

[31] Marc Hanussek, Matthias Blohm, and Maximilien Kintz. Can automl outperform humans? an evaluation on popular openml datasets using automl benchmark. *arXiv preprint arXiv:2009.01564*, 2020.

[32] R. Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 01 2000.

[33] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, ICSE-SEIP '19, page 291–300. IEEE Press, 2019.

[34] Stefan Studer, Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. Towards crisp-ml(q): A machine learning process model with quality assurance methodology. 03 2020.

[35] Eric Breck, Shanqing Cai, E. Nielsen, M. Salib, and D. Sculley. The ml test score: A rubric for ml production readiness and technical debt reduction. *2017 IEEE International Conference on Big Data (Big Data)*, pages 1123–1132, 2017.

[36] David Underdown et al. Quantifying digital preservation risks using statistics. https://blog.nationalarchives.gov.uk/quantifying-digital-preservation-risks-using-statistics/.

[37] Marlene Manoff. *Mapping archival silence: technology and the historical record*, page 63–82. Facet, 2016.

[38] Jenny Bunn. Working in contexts for which transparency is important. *Records Management Journal*, 2020.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[40] Sean Gillies et al. Shapely: manipulation and analysis of geometric objects, 2007–.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[42] Zhi Chen, Wu Guo, Li-Rong Dai, Zhenhua Ling, and J. Du. Neural text clustering with document-level attention based on dynamic soft labels. In *INTERSPEECH*, 2019.

A Natural Language Processing

A.1 Standard text data cleaning processes using NLP methods

Document de-duplication is the process of removing duplicates from the document set. There are two reasons for this step. Firstly, the model can become biased towards documents which the algorithm sees more than once. Secondly, there is a danger that the duplicate of a document could appear in both the training and test data which could over-inflate the accuracy score. However, due to the class representation in the corpus being highly imbalanced minor classes are made even smaller by de-duplication. The process was tested both with and without this step.

Document normalisation is the process of standardisation of text information. All documents are converted to .txt format to extract text data and metadata using Apache Tika. While standardisation is convenient for subsequent processing it does lead documents losing their original structure. For the algorithms used this was appropriate but if structure is considered important then a more sophisticated approach is needed.

Tokenisation is the task of breaking a sentence into smaller parts: single or multiple words. Several third party NLP libraries such as NLTK, Spacy and Gensim offer support functions to tokenise data. The tokenization step also removes common words (stopwords) such as ‘is’, or ‘was’ which can distort results. Additionally, punctuation was removed and all text was converted to lowercase. This conversion may hinder the contextual meaning of the data (particularly for proper nouns and acronyms), but it simplifies the computation if the text is standardised.

Lemmatisation is the process of grouping together the inflected forms of a word so that they can be analysed as a single item. For example, 'better' and 'good' have the same lemma, as do 'walk' and 'walking'. This has the effect of reducing the size of the vocabulary and the number of features which need to be learned, and therefore makes the most of the corpus. It means if a word appears once in each of three forms they can be pooled together as one feature. The downside is that the original meaning can be lost.

Stemming is a less sophisticated alternative to lemmatisation which removes suffixes from words. For example, 'fishing', 'fished', and 'fisher' are reduced to 'fish'.

Count Vectoriser converts a collection of text documents to a matrix of token counts and builds vocabulary from it. The matrix is generally sparse as any particular document will contain a small subset of all the words found in the corpus. If the corpus had a 5000 word vocabulary the document "cat do cat" would contain a 2 in the "cat" column and a 1 in the "dog" column, and zero in all other columns. Once a Count Vectoriser has been built from a corpus new documents can be encoded using the representation, but previously unseen words will be left out.

TF-IDF Transformer is similar to the count vectoriser but instead of word counts it weights words according to their frequency in the document (TF: Term Frequency) and the number of documents they appear in (IDF: Inverse Document Frequency). One advantage over simple word counts is that it weights more highly words which are important to a particular document while reducing the weighting of common words which are not necessarily stop words.

Clustering of different documents is based on the features we have generated using above procedures.

B Acronyms Quick Reference

Table 2: Acronyms

Acronym	Full form
AI	Artificial Intelligence
BI4BG	Better Information for Better Government
CLI	Command Line Interface
EDRMS	Electronic Document and Records Management System
GUI	Graphical User Interface
KIM	Knowledge and Information Management
ML	Machine Learning
NLP	Natural Language Processing
PRA	Public Records Act
RM(s)	Records Manager(s)
TFIDF	Term Frequency Inverse Document Frequency
TNA	The National Archives
XAI	Explainable AI