Market Research into AI/ML Tools For Document Selection and Classification

1. Introduction

The National Archives is the country's guardian of over 1000 years of iconic national documents. They are also the expert advisers in information and records management building on over 170 years of pioneering work in managing official public records. The National Archives is responsible for providing advice to government departments on their selection of records for transfer to the archive.

Today government departments are faced by a new challenge, one that has been described as a digital tsunami. A tsunami made of waves of born-digital records scattered across shared drives, and electronic document and records management systems (EDRMS), usually in an unstructured or semi-structured format. When considering what constitutes born-digital records, you start to realise the magnitude of this digital tsunami; which includes text-based documents, email, presentations, spreadsheets, images, CAD drawings, 3D models, data sets and data bases. If we now multiply these diverse formats of born-digital records with the huge amount of digital data being produced every second we start to see the true scale. The huge amounts of born-digital records make it nearly impossible for government departments to rely solely on manual techniques for selection and preservation.

> An attempt to ride the Tsunami wave rather than drown in an ocean of born-digital data

As data consumption and accumulation continues to grow so does the fields of data science, machine learning and artificial intelligence. The National Archives has always had a proximity to the world of AI/ML through its ongoing research, testing and partnerships.

The growth of AI and ML is happening at a very fast pace and the availability of commercial applications that are relatively easy to use are growing and evolving. Areas such as NLP, text classification, and document classification have been seeing wide improvements especially in industries such as legal and Finance, where several commercial tools exist to facilitate the extraction of key information from large amounts of legal documents and financial invoices.

Over the past few years the National Archives has been researching and testing different ML/AI applications to assess how Artificial intelligence can support and enhance the process of document classification and selection.

The aim of this report is to provide an in-depth review of the state of AI tools across the whole AI landscape, to identify tools and approaches that could be used to identify records suitable for the selection of born-digital records for transfer to The National Archives. The report also shortlists a few tools that meet the general functional requirements and then provides a recommendation to test four of these shortlisted tools.

The report also provides an overview of the proposed testing activities, benchmarking exercise and recommended statistical methodologies for evaluation.

2. Analytical Approach

The first step in our research is to really understand how document selection and classification works, and therefore breakdown the main requirements for a tool that uses AI to assist in the selection of records for preservation into smaller precise requirements. The approach allows us to understand the precise features and functionalities needed, and how we can use these features when assessing if a tool is fit for purpose. Additionally, what if we can't find the silver bullet, that magic tool that does everything? We can identify which tools are fit for which requirement thus allowing us to become more efficient and effective with parts of the process.

Several interviews have been conducted with relevant digital preservation stakeholders including digital archivists, information managers, digital preservation managers, project managers, data scientists and researchers. These interviews enabled the formulation of a clear set of requirements, which can be used as an assessment criteria against the tools being researched and can also help determine tools that can be shortlisted for additional research and testing.

A silver bullet might not exist! But can these tools make us more efficient?

Diagram 2.1: Research Steps

Table 2.1 shows the tool requirements collected during our research. We've also split the requirements into 'Must Have' and 'Nice to Have' features to ensure the most important requirements are prioritised and tools aren't being rejected for missing less critical features.

For a government department to go through their huge amounts of data and make critical decisions around preserving or destroying the contained content, they need to be able to classify the underlying documents or digital records and their content. Two categories of which are considered the most essential functional requirements, 'document classification' and 'text classification', have then been split down into individual requirements that are needed to facilitate this process of classifying records and their content.
In addition to these two main requirement groups others included security, cloud based, the type of licensing, ability to work with different file formats and ease of use.

A single tool might not be able to tick all the boxes, but we would still want to know the ones it ticks well. The challenge on hand is complex and different departments will have different priorities towards individual features

2.1 Requirements

REQUIREMENT	DETAILS	FEATURE IMPORTANCE
Document Classification. The tool can classify documents using	File Location	Must Have
	File Date	Must Have
	File Format	Must Have
	Document Name	Must Have
	Similarity between files within the same	Nice to Have
	Document size vs similar documents	Nice to Have
Text Analytics/Classification The ability to provide context	Identify Organisation	Must Have
	Identify Department	Must Have
	Identify People	Must Have
	Predict Topic	Must Have
	Predict document importance	Must Have
	Identify Title	Must Have
	Differentiate between internal and external people	Nice to Have
	Routine vs non-routine document	Nice to Have
Security	Access Levels	Must Have
Platform	Cloud	Must Have
License	Open Source	Nice to Have
Algorithms	Ability to provide some explanation to how the	Must Have
Machine Learning Method	Unsupervised Method	Nice to Have
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email	Must Have
File Locations	Access data from multiple cloud providers	Must Have
	Access data on shared drives	Must Have
UX	Easy to use front end interface	Nice to Have
	Allows collaboration between users	Nice to Have

3. The search begins

In a busy world of digital tools and the growing hype of buzz words like AI and ML, our research not only requires exploring as many tools as possible, it also requires trying to ensure these tools actually deliver on what they promise. Additionally, some tools might be very useful for certain industries or documents such as financial, legal or medical text, but not suitable for a wider scale implementation across government departments.

Our data collection and tools identification exercise involved using search engines, reviewing forums and online discussions, YouTube videos, published academic papers, online case studies, supplier websites, technology websites, foreign archiving departments and relevant organisations. We ensured our research was not limited to the UK and covered as much as possible from the global AI landscape.

The research lead us into reviewing 24 tools, technologies, products and models that might be relevant to our requirements. As a result, we dug deeper into each tool through publicly available documentation on suppliers websites, technical documentation, whitepapers or elsewhere on the Internet.

3.1 The state of AI/ML tools

Earlier in the research, it become clear how this market is becoming very busy and growing at unprecedented rates. These rates come at no surprise as the AI software market is expected to reach $118.6 Billion in annual worldwide revenue by 2025, according to market intelligence firm Tratica. Additionally, there is a shift from suppliers, businesses and organisations talking about AI and its benefits to actually implementing AI and publishing case studies, also success stories.

Most software and solution providers are tapping into the AI/ML revolution ranging from simple solutions that provide and add on AI functionality, to complete advanced enterprise grade record management systems with state-of-the-art Artificial intelligence and machine learning. Large enterprise companies are enriching their platforms with more advanced AI/ML capabilities, and also new startups that are entering the market are building their own branded solutions built on the technology of other large platforms and technology providers.

The market is also seeing continuous improvement in the field of Natural Language Processing, especially around the use of deep learning, which includes word embeddings and neural networks. This improvement is also accompanied by the increase in demand for NLP solutions by both the private and public sectors.

As the process of digging into each tool/technology was underway, we quickly realised that they can be split into three main categories:

A) Software Packages: readymade software solutions
B) AIaaS/MLaaS: Artificial Intelligence as a service / Machine learning as a service
C) Open Source: Open source technologies, code and algorithms

The below table 2.1 states the tools/technologies the research covered.

TOOLS THAT THE RESEARCH LOOKED INTO	
Open Text	Amazon Comprehend
Shiny Docs	Auto Keras
Netwrix	Auto-sklearn
Google Cloud Auto ML	ML Box
IBM Watson	TPOT
Keras and Tensor flow on Google Cloud ML	NLTK
Microsoft Azure ML	AYLIEN
Azure Cognitive Services	Meaning Cloud
Parascript	Adlib Elevate
Extract Systems	On Base By Hyland
Smart Soft - Document Classification SDK	Parallel Dots
Records 365 by RecordPoint	Alfresco

3.2 Shortlisting

The research then moved onto creating a short list of tools that are most relevant to addressing the requirements on hand. This shortlisting exercise relied on:

- Content available on the supplier's websites including case studies and documentation
- Tools ability to serve in a large public sector or enterprise scale environment
- Excluding tools that are designed to serve specific industries e.g. legal

This exercise lead to shortlisting nine tools/technologies. These tools have then been extensively investigated and assessed as much as possible through desk research without any actual testing. The nine shortlisted tools/technologies and a breakdown of our analysis are listed in the 'further research' section.

4. Further Research

4.1. Open Text

OpenText provide a suite of document and information management tools, with varying applications and scope. Three of these tools have key features that are able to meet some of the previously discussed requirements. These tools are Enterprise Content Management, AI & Analytics Suite and OpenText Discovery Suite.

OpenText Enterprise Content Management is a content and document management tool with in-built archiving and analytical features. OpenText offers a variety of features including ECM integration which supports applications such as SAP ERP, Oracle E-business Suite, Salesforce, Microsoft Office 365 and SharePoint. The platform also offers content archiving solutions allowing secure archiving, ECM collaboration and a simple user interface. This tool also offers flexible deployment options through private, public or hybrid cloud deployments.

One key relevant feature of Enterprise Content Management is the 'File Intelligence' implementation which uses a rules-based analysis on document metadata and keywords to provide a document archiving/deletion scheme. File Intelligence allows organisations to understand what information is inside their documents, including sensitive and regulated information such as PII and PCI. This feature also offers a multitude of extra capabilities including data source crawling, indexing, classification, search and reports tools and the implementation of specific actions such as deletion, tagging, archiving.

OpenText AI & Analytics Suite is an analytical tool with built in dashboard functionality. One feature, OpenText Magellan, has a relevant application for this project as it contains text analytics and predictive modelling capabilities. However, it is unclear how any NLP and textual analysis methods are integrated in their document management pipeline.

The following tool shows the available techniques, which include topic summaries and entity detection and with an example provided in http://magellan-text-mining.opentext.com/

OpenText Discovery Suite is a full end-to-end analytics tool using AI and ML to automatedly provide insights into document content. Although, it is unclear what classification tools are available with this product.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Yes: OpenText File Intelligence can classify (using a rules-based system) documents using metadata
	File Date	Yes: OpenText File Intelligence can classify (using a rules-based system) documents using metadata
	File Format	Yes: OpenText File Intelligence can classify (using a rules-based system) documents using metadata
	Document Name	Yes: OpenText File Intelligence can classify (using a rules-based system) documents using metadata
	Similarity between files within the same subdirectory/folder	No: OpenText uses rules/methods that work on a document-by-document basis, independently of other files
	Document size vs similar documents	No: OpenText uses rules/methods that work on a document-by-document basis, independently of other files
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes: OpenText AI & Analytics supports Entity recognition
	Identify Department	Yes: OpenText AI & Analytics supports Entity recognition
	Identify People	Unknown
	Predict Topic	Yes
	Predict document importance	Unknown: It is not clear how predictive modelling functions in OpenText AI & Analytics, and to what extent its features are capable.
	Identify Title	Yes: OpenText AI & Analytics supports Entity recognition
	Differentiate between internal and external people	Unknown: It is not clear entities in can be compared with a pre-set list of internal people
	Routine vs non-routine document	Unknown: It is not clear how predictive modelling functions in OpenText AI & Analytics, and to what extent its features are capable.

Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No
Algorithms	Ability to provide some explanation to how the selection process works	Algorithms are either rules-based or undocumented.
Machine Learning Method	Unsupervised Method	Algorithms are either rules-based or undocumented, it is not clear what type of ML method the Predictive Modelling Tool OpenText AI & Analytics uses.
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	MSG is unknown
File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Yes

4.2. Adlib Elevate

Adlib Adlib is a document management, archiving and conversion tool with additional NLP and document classification capabilities as part of their Adlib Elevate, File Analytics functionality and Progressive Document Classification Solution. A key component of these services is the detection of what Adlib refers to as ROT (redundant, obsolete, trivial). This matches precisely a key requirement in digital archiving, allowing the detection and then removal of unimportant documents.

Progressive Document Classification

Adlib's Progressive Document Classification uses unsupervised machine learning methods, including clustering, to classify and organise documents. Pre-processing carried out in Adlib Elevate will organise text and document content into a unique file "fingerprint", which is then used in a clustering process to provide document categories. Semi-supervised (rule building) methods can also be implemented to manually create automated classification (and categorisation) pipelines.

File Analytics

Adlib's File Analytics is a collection of document and text analytics tools, allowing the detection of ROT documents and their management/deletion. The precise method (including to what length different document properties are considered, e.g. metadata, other files' content) of this detection step is undisclosed, along with requirements (how many documents are needed before it can accurately pick out e.g. obsolete/redundant elements) and accuracy.

Integration with data sources:

Adlib Elevate integrates with a range of sources including Microsoft SharePoint, Office365, Shared Drives, and OpenText.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Yes
	File Date	Yes
	File Format	Yes
	Document Name	Yes
	Similarity between files within the same subdirectory/folder	Unknown
	Document size vs similar documents	Unknown
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes
	Identify Department	Yes
	Identify People	No
	Predict Topic	Yes
	Predict document importance	Yes. ROT detection will pick out unimportant documents.
	Identify Title	Yes
	Differentiate between internal and external people	No
	Routine vs non-routine document	Yes. ROT detection will pick out unimportant documents.
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No
Algorithms	Ability to provide some explanation to how the selection process works	Unknown
Machine Learning Method	Unsupervised Method	Has both supervised and unsupervised capabilities.
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Yes
File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Yes

4.3. Record 365

Records365 is a Cloud-based content storage and information management tool developed by RecordPoint. Its purpose is to provide end-users with an intelligence driven management system, which is able to connect and integrate with a multitude of different platforms and data sources, including SharePoint and Office 365. Unlike many other NLP tools, Records365's primary task and specialism is towards document management and analysis, as opposed to more general cloud service or software tools packaged with additional language or text analytics features.

Classification Intelligence and Rules Based Categorisation

A recent addition to Records365 is the ability to train by category ML classification models (a supervised method which will require pre-labelled data), that can then be used to further predict and classify categories for future documents. This is a complete code-free process and can be used in conjunction with non-ML rules-based approaches. Only the content of a document is used in Records365 in its classification methods.

Metadata from the file can also be included in a larger model by using the pre-existing classification functionality of the Records365 tool, by decision based-rules. These rules are easily created, tree-based decisions that apply a manually created step-by-step algorithm to each file and can be included as a separate combined stage with the previous classification routine. This approach is capable of scaling to arbitrarily large datasets, and requires no training-stages for the rules-based step.

Primarily the software is oriented around document management, and therefore has the ability to integrate document classification with its more general archiving features. For example, after applying a rules-based system or classification method to our datasets, we can then easily apply a file deletion routine to certain categories or files satisfying specific criteria.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Yes
	File Date	Yes
	File Format	Yes
	Document Name	Yes
	Similarity between files within the same subdirectory/folder	No: each file is considered individually
	Document size vs similar documents	No: each file is considered individually
Text Analytics/Classification The ability to provide context	Identify Organisation	Recognition of entities is unavailable, but a rules-based algorithm can be manually developed to single out specific named organisations.
	Identify Department	Recognition of entities is unavailable, but a rules-based algorithm can be manually developed to single out specific named organisations.
	Identify People	Recognition of entities is unavailable, but a rules-based algorithm can be manually developed to single out specific named organisations.
	Predict Topic	No: Topic modelling is unavailable in Records365, as are Key Phrase identifying capabilities.
	Predict document importance	No: Classification is carried out by assigning documents into *categories*. It may be possible to set up a rules-based system which can decide whether a file is important/unimportant, but the logic of this algorithm would be human made, rather than an ML/AI solution.
	Identify Title	Yes
	Differentiate between internal and external people	No: Entity Recognition is unavailable, and it will not be possible to create a rules-based system that can pick out a person, unless that person's name is given beforehand.
	Routine vs non-routine document	Possibly – Although Supervised and arbitrarily accurate: Records365 Classification Intelligence

		could be used to classify documents into routine/ non-routine. This will require a training set with labelled classes.
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No
Algorithms	Ability to provide some explanation to how the selection process works	Rules based methods are created by users and is easy to understand/explain through visual tools, whereas there is no documentation available for RecordPoint's ML methods.
Machine Learning Method	Unsupervised Method	Minor unsupervised learning
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Yes
File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Yes

4.4. Alfresco

Alfresco offers two services that provide the necessary tools for content and document management, these are Alfresco Content Services and Alfresco Governance Services. Alfresco Content Services has AI and ML functionality by using AWS Amazon Comprehend and a Search and Insight Engine implementation using Apache Solr. Alfresco Governance Services provides automated records management, allowing the detection and archiving or deletion of documents using traditional metadata and rules-based techniques.

Alfresco Content Services

Alfresco Content Services provides AI and intelligence analytics capabilities to its document management routine. As mentioned above, AWS provides the ML and NLP in Alfresco's Intelligence Service feature of Alfresco Content Services. These provide easy to use document recognition capabilities not available in other rules-based Alfresco products. No supervised methods are used (or available), only rule-based decisions need to be created to decide whether any intelligence services are to be applied.

Results are saved as file metadata which can be used as part of a rule-based system available in other Alfresco services, such as Alfresco Information Governance, which is available in Alfresco Content Services and Alfresco Governance Services.

Alfresco Governance Services

Primarily as a compliance or security tool, Alfresco Governance Services uses metadata and user-created rules to archive and dispose of documents. These rules can be used to create a categorisation, by determining where the record should be filed.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Yes: using Alfresco Governance
	File Date	Yes: using Alfresco Governance
	File Format	Yes: using Alfresco Governance
	Document Name	Yes: using Alfresco Governance
	Similarity between files within the same subdirectory/folder	No
	Document size vs similar documents	No
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes
	Identify Department	Yes
	Identify People	No
	Predict Topic	No
	Predict document importance	Yes
	Identify Title	Yes
	Differentiate between internal and external people	No
	Routine vs non-routine document	Yes: using Alfresco Governance
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No
Algorithms	Ability to provide some explanation to how the selection process works	NLP/ML is carried out through AWS and will depend on AWS documentation
Machine Learning Method	Unsupervised Method	Yes, using Amazon Comprehend
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Yes
File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Unknown

4.5. Google Cloud

Google Cloud Platform hosts several AI/ML services which can be used to build a document classification and archiving tool. Two of these can be used in this application: Google Cloud AutoML and Google Cloud Natural Language, which together can provide an end to end document management system.

Google Cloud AutoML provides a small set of ML tools that can be trained on data stored on Google Cloud Platform.

Google Cloud Natural Language provides a set of pre-trained NLP functions. However, it will require further programming and development to fully integrate it into an overall project. The Natural Language API provides the functionality for this to be used as part of an application.

Google Cloud AutoML relevant capabilities

- Custom entity extraction. Supervised learning to identify pre-set departments, organisations, and people.
- Custom content classification. Supervised learning to identify pre-set content classes.
- Integrated REST API
- Custom Sentiment Analysis
- Custom Models

Google Cloud Natural Language API relevant capabilities:

- Entity analysis.
- Content classification. A pre-trained model that can categorise a body of text into one of 700+ categories. Many of these are very relevant to our case with a range of very specific categories and subcategories.
- Syntax analysis
- REST API's
- Sentiment Analysis

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	No: Uses text body only
	File Date	No: Uses text body only
	File Format	No: Uses text body only
	Document Name	No: Uses text body only
	Similarity between files within the same subdirectory/folder	No: Works on an individual file basis.
	Document size vs similar documents	No: Uses text body only
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes
	Identify Department	Yes
	Identify People	Yes
	Predict Topic	Yes
	Predict document importance	Unknown
	Identify Title	No
	Differentiate between internal and external people	No
	Routine vs non-routine document	No
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No, but allows running open source technology
Algorithms	Ability to provide some explanation to how the selection process works	Unknown
Machine Learning Method	Unsupervised Method	Both supervised and unsupervised
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Unknown
File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Yes

4.6 Microsoft Azure

Microsoft Azure provides two Machine Learning Frameworks, Machine Learning Studio and Azure Machine Learning Service, which allow users to create a full end-to-end ML experience. Azure Machine Learning Studio is a graphical drag-and-drop interface allowing a user to create a full ML pipeline using in-built pre-processing and ML models (Modules). Whereas Azure Machine Learning Service allows users to build, train and develop their own ML models, using Python, R and a selection of popular Azure modules.

Relevant NLP and other modules include:

- Machine Learning Studio LDA. (This module is similar to Topic Summary, but no textual information about the topics is outputted, just a scoring to be used as part of a larger model).
- Machine Learning Studio Extract Key Phrases from Text
- Machine Learning Studio Named Entity Recognition
- Machine Learning Studio Clustering
- Machine Learning Studio Anomaly Detection (classification and so requires labelled data)

Other Azure products of note that can be considered include Microsoft Azure Cognitive Services. This provides:

- Cognitive Services Text Analytics Named Entity Recognition
- Cognitive Services Text Analytics Key Phrase Extraction
- Cognitive Services Text Analytics Forms Recogniser

Classification model creation is also provided as an ML tool using one of multiple model initialisations. However, as with all classification platforms or models or techniques, their training will require pre-labelled datasets. If pre-labelled datasets (large) are provided, then Azure's classifications models can be used directly. Otherwise a separate binary or probabilistic method might need to be developed, through Azure Machine Learning Service, which could be included to predict document importance, document relevancy, topic, title and (if possible) file metadata.

As Azure Machine Learning Studio does not support the automated training of models, each step involving a separate clustering, or LDA (Latent Dirichlet Allocation), also other models (for example checking how similar the contents are of a new folder or sub-folder) will need to be manually retrained when moving to a different dataset. Each one of these will require manually running the training on the interface. This will be necessary when looking at in-folder similarity and importance across completely new topics.

Azure has dataset size limitations when selecting modules, which means that large projects may have to be carried out in chunks.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Possible using Azure Machine Learning studio with Python/R scripts
	File Date	Possible using Azure Machine Learning studio with Python/R scripts
	File Format	Possible using Azure Machine Learning studio with Python/R scripts
	Document Name	Unknown: Could require Python pre-processing
	Similarity between files within the same subdirectory/folder	Yes: Clustering and LDA. Limited automation. Each subdirectory/folder batch will need to be trained manually
	Document size vs similar documents	Unknown: Could require Python pre-processing
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes
	Identify Department	Yes
	Identify People	Yes
	Predict Topic	No: requires a custom ML model to create
	Predict document importance	No: requires a custom ML model to create
	Identify Title	Might need Python Script; ML Extract Key Phrases
	Differentiate between internal and external people	ML Named Entity Recognition; Cognitive Services Named Entity Recognition; ML Python Modules.
	Routine vs non-routine document	Yes, Might require labelled data or Python Script; ML LDA; ML Clustering
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No, but allows running open source technology
Algorithms	Ability to provide some explanation to how the selection process works	Unknown
Machine Learning Method	Unsupervised Method	Yes. A mix of pre-trained models, supervised and unsupervised learning
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Yes. MSG Unknown

File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Unknown

4.7 Amazon AWS

Amazon Comprehend, part of Amazon's larger AWS platform, is an automated NLP service built to provide insights from textual documents. This service provides continuously trained and updated methods for multiple textual insights, including:

- Entity Recognition, characterising named entities found by entity and type e.g. Organisation, people. In addition Custom Entity Recognition exists, allowing provided concepts external to AWS to be recognised.
- Key Phrase Extraction, in which provides a list of individual insights and their occurrences from a single document.
- Topic Modelling, allowing a set of topics to be found from a collection of documents. A scoring or relevancy is then provided for each topic or document pair. In Amazon Comprehend, the keywords of these topics are explicitly given, rather than the topics only being numbered.
- Custom Classification, which although it requires labelling and may not be relevant for this project, is a supervised version of exactly what National Archive are looking for. Given a set of (labelled) documents, AWS will train a document classification model that can be used to classify all further documents. TThis classification can be multiple classes (e.g. minutes/non-routine/unknown/unimportant…), or a simple important or unimportant classification. Each classification result comes with an associated confidence or probability therefore allowing cut-off thresholds to be established.

All insights delivered through AWS comprehend must come from the text contents itself. Aspects such as metadata, even file type, amongst others, play no part in the ML models. Any such information will have to be provided to a separate model to account for these.

On its own, Amazon Comprehend does not have the ability to build an unsupervised model that will determine importance or unimportance of documents. It can, however, provide both a pre-processing step as part of a larger model using the results of Topic Modelling.

Amazon SageMaker

Amazon Comprehend alone will not be enough to match the model requirements, including the main goal of predicting document importance. The solution here will be to use Amazon SageMaker, Amazon's ML service to build predictive models, using both the text files and the results of Amazon Comprehend. These will be custom models and require an understanding of Python and ML/AI to complete.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Unknown: depends on metadata availability once file is uploaded into S3
	File Date	Unknown: depends on metadata availability once file is uploaded into S3
	File Format	Unknown: depends on metadata availability once file is uploaded into S3
	Document Name	Yes: using SageMaker
	Similarity between files within the same subdirectory/folder	Yes: Amazon Comprehend Topic Modelling
	Document size vs similar documents	Yes: using SageMaker
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes
	Identify Department	Yes
	Identify People	Yes
	Predict Topic	Yes
	Predict document importance	Yes: using SageMaker
	Identify Title	No
	Differentiate between internal and external people	Yes: using SageMaker
	Routine vs non-routine document	Yes: using SageMaker
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No, but allows running open source technology
Algorithms	Ability to provide some explanation to how the selection process works	Unknown
Machine Learning Method	Unsupervised Method	Both supervised and unsupervised
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Yes. Images might require Amazon Rekognition
File Locations	Access data from multiple cloud providers	Yes

	Access data on shared drives	Yes
UX	Easy to use front end interface	Yes
	Allows collaboration between users	Unknown

4.8 IBM Watson

IBM Watson has two NLP features available for our current requirements: Natural Language Understanding and Natural Language Classification, which can be combined with IBM's Deep Learning and other ML framework tools. The APIs available can then be used to develop a full-end-to end document classification tool.

Natural Language Understanding is a selection of NLP features including sentiment analysis and entity recognition. These are pre-trained models which, for our application, can be used to identify keywords, concepts and entities.

Natural Language Classification allows text-based classification, with some limitations. There is a 2048 character (300-400 word) per document limit and a requirement to use an API. The input of this API is a text string, not a document, and so a further programming and data processing step will be necessary. Only 30 text inputs can be classified per API request.

IBM Watson Studio

IBM Watson Studio is a no-code graphical tool that can be used to auto-generate ML models, with the ability to also integrate Python, R and Scala. The ML tools available in IBM Watson Studio includes a Natural Language Classifier model builder, and access to further IBM Watson tools using the API in notebooks.

Data to be used in IBM Watson Studio projects can either be from some IBM Cloud sources or a selection of third-party services. Any text will still need to be extracted from these files in a prior pre-processing step, and the Natural Language Classifier tool in IBM Watson Studio requires the data to be in CSV format, with assigned labels for model training.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	No: Metadata can only be used for URL and HTML inputs
	File Date	No: Metadata can only be used for URL and HTML inputs
	File Format	No: Metadata can only be used for URL and HTML inputs
	Document Name	No: Metadata can only be used for URL and HTML inputs
	Similarity between files within the same subdirectory/folder	No: Only the text itself is used.
	Document size vs similar documents	No: Metadata can only be used for URL and HTML inputs
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes
	Identify Department	Yes
	Identify People	Yes
	Predict Topic	Yes
	Predict document importance	Possibly Yes
	Identify Title	No
	Differentiate between internal and external people	Yes, but will require knowledge of internal individuals: Uses Custom Models and Entities Recognition in Natural Language Understanding
	Routine vs non-routine document	Possibly Yes
Security	Access Levels	Yes
Platform	Cloud	Yes
License	Open Source	No, but allows running open source technology
Algorithms	Ability to provide some explanation to how the selection process works	Unknown
Machine Learning Method	Unsupervised Method	Both supervised and unsupervised

File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	No, requires text pre-processing and works with CSV files only
File Locations	Access data from multiple cloud providers	Yes
	Access data on shared drives	Yes
UX	Easy to use front end interface	Unknown
	Allows collaboration between users	Unknown

4.9 Open Source: Python Libraries

Gensim: Topic modelling for humans

Gensim (generate similar) is a python library used for multiple NLP applications and is built to handle large datasets. Word embeddings can be trained, or pre-trained models can be imported separately. This library contains word2vec and doc2vec implementations with ML algorithms including LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation).

This library contains topic summary implementations that can be used to categorise collections of documents or provide summaries of individual documents. Similarity implementations also exist, which can be used to compare documents.

Topic summary implementations will be able to identify key terms, organisations and departments as well as topics from a collection of documents. Document importance can be established from the depth of detail or keywords of the summary. Importance can be established using similarity models to detect redundant documents.

SpaCy: Industrial-Strength Natural Language Processing

Spacy is an NLP library designed to be used in production applications. It contains a library of high performance pre-trained and unsupervised word embeddings and NLP models. SpaCy also integrates with deep learning frameworks including TensorFlow, Keras and Sci-Kit Learn, and has entity identification functionality. Integration with these frameworks is necessary to provide a complete classification routine.

The libraries pre-trained word embeddings can be used as part of a pre-processing stage of a classification pipeline. The Entity Identification tool will allow the detection of organisations, departments and people in documents.

REQUIREMENT	DETAILS	OUR RESEARCH
Document Classification. The tool can classify documents using	File Location	Yes: usingpython libraries
	File Date	Yes: using python libraries
	File Format	Yes: using python libraries
	Document Name	Yes: using python libraries
	Similarity between files within the same subdirectory/folder	Yes: using Gensim
	Document size vs similar documents	Yes: using python libraries
Text Analytics/Classification The ability to provide context	Identify Organisation	Yes: using Spacy
	Identify Department	Yes: using Spacy
	Identify People	Yes: using Spacy
	Predict Topic	Yes: using Gensim
	Predict document importance	Yes: using Gensim
	Identify Title	Possible
	Differentiate between internal and external people	Unknown

	Routine vs non-routine document	Unknown
Security	Access Levels	No, unless run on a MLaaS
Platform	Cloud	Yes
License	Open Source	Yes
Algorithms	Ability to provide some explanation to how the selection process works	Yes
Machine Learning Method	Unsupervised Method	Both supervised and unsupervised
File Formats	WORD, PDF, EXCEL,Text, Email (MSG), Email Attachments and Images	Yes
File Locations	Access data from multiple cloud providers	Requires using a cloud provider
	Access data on shared drives	Requires using a cloud provider
UX	Easy to use front end interface	No
	Allows collaboration between users	No, unless running on a cloud solution that does

5.0 Recommended for Testing

Desk research is never conclusive enough to decide if any of the reviewed tools can deliver the desired outcome. However it won't be feasible to test every tool researched. Therefore, the research developed a recommended list of four tools to be explored further and tested using the testing process detailed in section six of this report.

This final selection of tools to be tested considered the below points when a decision was made:

Individual Requirements

How the tool performs against each individual requirement

Overall Outcome

The possibility of the tool achieving the overall outcome

Ecosystem

How the tool works with the wider pipeline/ecosystem

Integration

The tool's ability to integrate with other commonly used government technologies, tools or platforms.

Time Limitation

The ability to test the tool in a short timeframe with minimal setup

Diversity

Ensuring tools from different groups are tested. i.e. Software Packages vs AIaaS/ MLaaS

Four Recommended Tools for Testing

- Records365
- Adlib Elevate
- AWS Comprehend
- Azure ML

Additionally, a recommendation has been made to use open source libraries or models as a benchmark during the testing phase against the shortlisted tools.

6.0 Testing

It's now time to put our four recommended tools to the test and assess how well these tools deliver against the requirements and the complexity involved in achieving the desired outcome. The aim of this testing phase is not to announce a winning tool, but rather to further understand which tools are good at doing what and how can these tools reduce the need for manual classification and selection.

Considering the testing requirements, its clear that we have two separate testing needs. The first is a more of a software functional testing approach to assess the software ability to deliver against our individual requirements. The second test is around the tools ability to deliver the final desired outcome and its machine learning and artificial intelligence accuracy and error rates. Accordingly, the recommended testing has been split into two activities; modular testing and outcome-based testing.

These tests will be conducted by the tool's suppliers, partners or qualified testers to ensure each tool's maximum potential is unleashed.

> We are not looking for one winning tool. We are evaluating which tool is good at what and how can they help us become more efficient

6.1 Modular Testing

Inspired by the modular testing framework used in automated software testing, a software is split into different units or functions and tested in isolation. Our research recommends using a similar approach, aiming to evaluate the functionality of each tool against an individual requirement in isolation from the other. E.g. can the tool detect the file date?

This not only helps understanding the features offered by each tool, but also helps government departments understand which tools excel at which individual features, enabling themselves to make more informed decisions based on their individual requirements.

The test will be conducted manually, and scores are allocated based on the following marking scheme, as shown in diagram 6.1. The scoring considers two critical factors:

- Can the tool achieve an individual requirement being tested?
- In order to achieve this requirement, does the tool require any additional customisations or manual coding?

It is important to understand not only if a tool can can deliver against the requirements, but also if customisation or coding is needed, and the complexity of these changes. Again this will be beneficial for government departments to assess the complexity of using one of these tools to achieve a specific outcome.

Diagram 6.1 Module Based Testing Flow

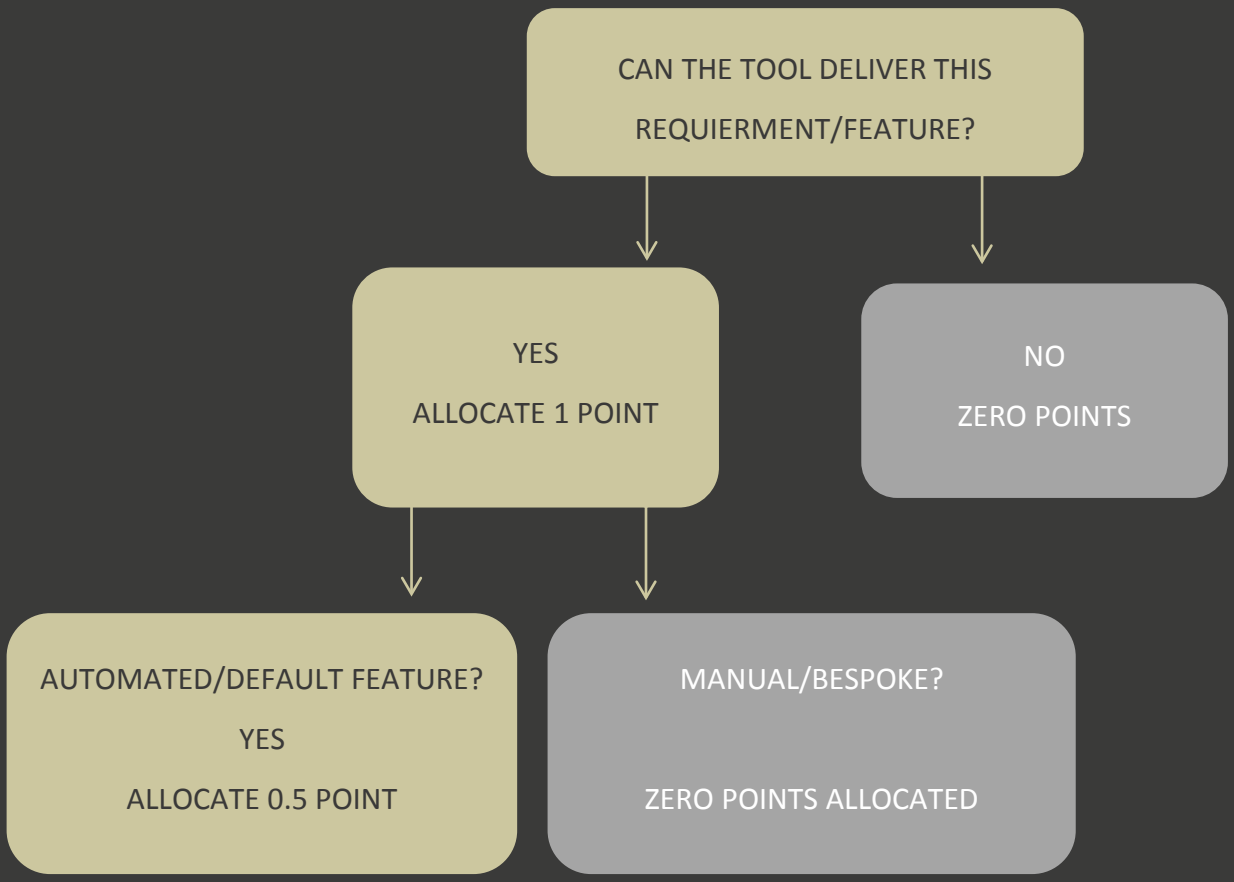


Figure 6.2 Preview from Testing Template

Category	Feature	YES/NO	Automated/ Manual
Document Classification The tool can classify documents using	Detect file location		
	Detect file date		
	Detect file format		
	Detect document name		
	Detect and relate document size vs similar documents		
	Detect similarity between files in the same folder/subdirectory		
Document Classification Total	Total Score		
Text Classification/Analytics The tool can provide context	Identify Organisation		
	Identify Department		
	Identify People		
	Predict Document Importance		
	Identify Title		
	Predict Topic		
	Diffrentiate between internal and external users		
	Routine vs non-routine document		
Text Classification Total	Total of "yes"		

6.2 Outcome Based Testing

Once we've understood how these tools are performing against the list of requirements, it is important to understand how they perform against our main objective and deliver the desired outcome. The objective now is to understand if any of these tools can classify and select documents and decide which documents should be destroyed or archived. In doing so, what level of accuracy can be achieved and how risky is it to rely on these tools?

These questions lead to designing an additional outcome-based test to measure the tools ability to deliver on the final objective. In order to determine if a tool is accurately classifying and selecting documents, there is a need to provide three type of datasets.

Did a tool achieve a 95% accuracy, yet delete the most important documents?

Training Dataset:

This is the actual dataset to be used in training the model.

Validation Dataset:

This dataset is used to fine tune the model and is sometimes referred to as the 'dev set'.

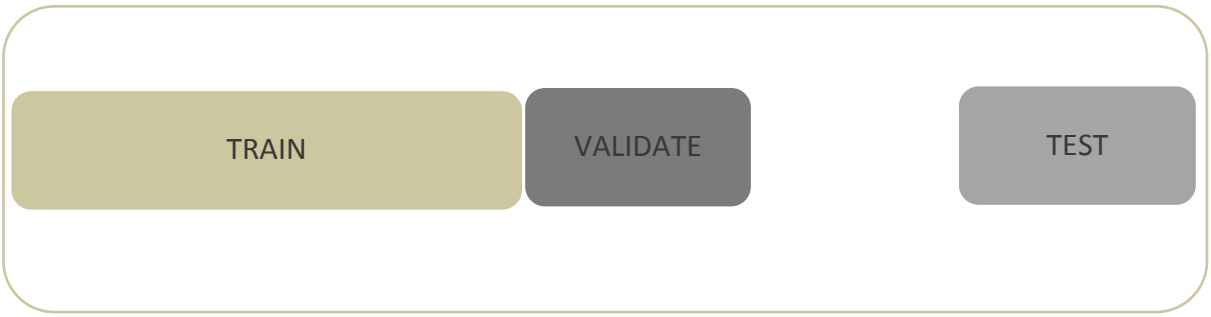
Test Dataset:

This dataset provides the outcome and evaluation of the model. It should only be used once the model is completely trained.

In order to achieve the best test datasets possible, we would need to consider the following:

- Using a sampling methodology when selecting the documents to be manually classified while ensuring the sample is representative of the different topics and file types available in the overall data set.
- The selected sample will then need to be classified and labeled as either destroy or preserve.
- Manually classify as many documents as possible, as the larger the sample size the more accurate the results.

Figure 6.3: Split between datasets



6.2 Testing Risks

As with any software evaluation and testing activity there are a few risks that need to be considered. Looking into our proposed testing frameworks, testing timeliness and restrictions, we've identified the below potential risks:

- Most machine learning testing and competition activities involve using large test and validation sets. The size of these datasets are usually in the hundreds of thousands, if not millions. Our test and validation sets are expected to be in the thousands.
- Small datasets can also make it more challenging when applying a statistical methodology for comparison.
- The individual requirements-based testing would require a dedicated internal resource to work alongside suppliers for assessment and scoring.
- The difference in the nature of the tools being assessed and how ready-made software packages work vs. MLaaS/AIaaS platforms. These include the ability to easily implement bespoke models and algorithms which isn't usually the case with packaged software solutions.

6.3 Methodology for Comparison

Although these tools aren't competing for a golden medal, it's still important to understand how they perform and how they compare to one another. This is straight forward in the case of the modular based testing approach against individual requirements, but not as straight forward with outcome-based testing.

The ability to accurately predict a document is important and the higher the success rate the better. Additionally, error rates have significant importance to the success of any of the tools being tested as it won't be possible to rely on a tool that has a high success rates, yet the ones it incorrectly classifies are our most important documents.

In order to achieve a better understanding of accuracy and error rates, and how these tools compare to one another, the proposed approach is the utilisation of the confusion matrix or the matching matrix in case of unsupervised learning.

The confusion matrix:

This is a table that is often used to describe the performance of a classification model or a classifier on a set of test data for which the true values are known. Table 6.3 below shows an example of a confusion matrix.

Table 6.3: Confusion Matrix.

	Predicted NO	**Predicted YES**
Actual: NO (e.g. no don't delete)	True Negative	False Positive
Actual: YES (e.g. yes delete)	False Negative	True Positive

Before considering how accuracy and misclassification rates are calculated, the below list provides a brief explanation of the different measures involved:

- **True Positive:** when a model correctly predicts the positive class. e.g. if a tool predicts 'yes, delete' and the file should be deleted.
- **True Negative:** when a model correctly predicts the negative class. e.g. if a tool predicts 'no. don't delete' and the file shouldn't be deleted.
- **False Positive:** when a model incorrectly predicts the positive class. e.g. if a tool predicts 'yes, delete' while the file shouldn't be deleted.
- False Negative: when a model incorrectly predicts the negative class. e.g. if a tool predicted 'no, don't delete' while the file should have been deleted.

Accuracy and misclassification rates:

- Accuracy: How often is the model correct? = (True Positive + True Negative)/Total
- Misclassification rate: How often is the model wrong? = (False Positive + False Negative) / Total

In addition to the above measures the comparison should consider using Null Error Rate, Cohen's Kappa, F-Score and ROC curve.

7.0 Conclusion

Looking at the nature of born-digital data and the complexity of the selection process, it is clear that finding one tool that ticks all the boxes proved difficult. However, this does not suggest that the tools available on the market today cannot dramatically improve the process and shift the human selection role into a quality assurance and audit role

The variety and diversity of the tools and technologies encountered during the research clearly demonstrate the growing demand for solutions that enable organisations to tap into their unstructured data repositories, and understand the nature and importance of this data. Not only for archiving purposes but also for more wider needs across industries, ensuring ongoing growth in the marketplace.

As previously discussed, the shortlisting and recommendation of tools does not intend to name one winning tool and disqualify the rest. The research aims to provide government departments with a wider understanding of the tools available and how they perform against specific requirements, thus enabling them to make informed decisions during any potential market engagements.

The testing phase will take this research one step further through enriching the report findings and will elaborate on how these tools perform when tested on real records and data sets. Additionally, the testing will uncover any potential challenges with certain tools when integrated with the wider data pipeline and ecosystem. This testing might also lead to a conclusion that some of the tools reviewed might actually work well together and complement one another rather than compete against each other. The testing will also provide solid evidence of how these tools performed towards the final outcome of recommending records to preserve or destroy using samples of real unstructured records similar to what would be expected across most government departments.

The report attempted to ensure a statistical methodology is used when comparing and evaluating tools, and ensured it looks into commonly used industry practices around testing and scoring. This resulted in recommending the utilisation of a confusion matrix and identifying the accuracy and misclassification rates. In the case of document selection and classification, misclassification rates might prove to be more important than other industries or use cases due to the risks associated with misclassifying records to be destroyed, leading to the loss of important records

The market is growing and changing at lightning speed, speed, for example during the six weeks it took to complete this research and report two of the tools researched have already announced new additional features to their solutions. This will continue to be the case as the market continues to grow and the tools offered that will continue to develop and evolve. This also suggests that frequent research and evaluation of AI tools available in the market will be required, to ensure government departments are up to date with the latest offering around archiving and document selection tools, powered by Artificial Intelligence and Data Science.

Finally, the research also suggests that the human role will continue to be needed to provide quality assurance, the necessary level of confidence and risk management while AI/ML tools continue to evolve and mature.

8. References

- https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7
- https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative
- https://dfordatascience.wordpress.com/2018/11/14/machine-learning-basics-confusion-matrics-error-matrics/
- https://towardsdatascience.com/receiver-operating-characteristic-curves-demystified-in-python-bd531a4364d0
- https://smartbear.com/learn/automated-testing/software-testing-methodologies/
- https://resultspositive.com/12-key-performance-indicators-for-qa-test-managers/
- https://www.tractica.com/about/overview/
- https://www.tractica.com/newsroom/press-releases/artificial-intelligence-software-market-to-reach-118-6-billion-in-annual-worldwide-revenue-by-2025/
- https://towardsdatascience.com/major-trends-in-nlp-a-review-of-20-years-of-acl-research-56f5520d473
- https://www.opentext.co.uk/file_source/OpenText/en_US/PDF/opentext-auto-classification-product-overview.pdf
- https://www.opentext.co.uk/products-and-solutions/products/discovery/auto-classification
- https://www.shinydocs.com
- https://cloud.google.com/automl/
- https://cloud.google.com/blog/products/gcp/problem-solving-with-ml-automatic-document-classification
- https://cloud.google.com/document-understanding/docs/#classify
- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/text-analytics
- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-classification
- https://aws.amazon.com/comprehend/
- http://www.nltk.org/
- https://aylien.com/text-analysis-platform/
- https://www.meaningcloud.com/
- https://towardsdatascience.com/bert-text-classification-in-3-lines-of-code-using-keras-264db7e7a358
- https://www.parascript.com
- https://www.recordpoint.com/site/2019/05/AHRC-Case-Study_Letter.pdf
- https://www.recordpoint.com/solutions/classification-intelligence/
- https://www.paralleldots.com/text-analysis-apis#intent
- https://www.youtube.com/watch?v=dEK8uAC_9Kc
- https://elischolar.library.yale.edu/cgi/viewcontent.cgi?article=1035&context=jcas
- https://futureproof.records.nsw.gov.au/machine-learning-and-records-management/
- https://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf
- http://www.ijdc.net/index.php/ijdc/article/view/162/230
- https://gallery.azure.ai/browse?s=document%20classification
- https://www.todaysoftmag.com/article/2657/automatic-classification-of-documents-using-natural-language-processing
- https://blogs.technet.microsoft.com/machinelearning/2017/02/13/cloud-scale-text-classification-with-convolutional-neural-networks-on-microsoft-azure/
- https://towardsdatascience.com/choosing-between-tensorflow-keras-bigquery-ml-and-automl-natural-language-for-text-classification-6b1c9fc21013
- https://medium.com/mlrecipies/document-classification-using-machine-learning-f1dfb1171935
- https://github.com/liu-nlper/DocumentClassification
- https://www.kaggle.com/c/ntucsie-wm2018-topic-modeling/notebooks
- https://www.cs.sfu.ca/~jpei/publications/EmailMining-KAIS.pdf
- https://towardsdatascience.com/how-i-used-machine-learning-to-classify-emails-and-turn-them-into-insights-efed37c1e66
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220976
- https://realpython.com/python-keras-text-classification/
- https://arxiv.org/pdf/1909.08402.pdf
- https://medium.com/mlrecipies/document-classification-using-machine-learning-f1dfb1171935

- https://www.todaysoftmag.com/article/2657/automatic-classification-of-documents-using-natural-language-processing
- https://opennlp.apache.org/
- https://cloud.google.com/blog/products/gcp/intro-to-text-classification-with-keras-automatically-tagging-stack-overflow-posts
- https://www.expertsystem.com/document-classification-works/
- https://www.idm.net.au/article/0012033-artificial-intelligence-records-management
- https://datasemantics.co/how-automated-document-classification-helped-save-tons-of-time-for-an-electronics-giant/
- https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f
- https://dftdigital.blog.gov.uk/2018/04/09/the-write-stuff-how-we-used-ai-to-help-us-handle-correspondence/
- https://www.paralleldots.com/text-analysis-apis#intent
- https://uk.mathworks.com/help/deeplearning/ref/alexnet.html
- https://mapr.com/resources/videos/document-classification-apache-spark/
- https://www.youtube.com/watch?v=2IjIMCJbLhc
- https://www.abbyy.com
- https://axis-ai.com/
- https://www.globema.com/machine-learning-data-classification/
- https://github.com/liu-nlper/DocumentClassification
- https://sflscientific.com/solutions/case-studies/case-study-document-classification
- https://arxiv.org/search/cs?query=document+classification&searchtype=all&abstracts=show&order=-announced_date_first&size=50
- https://arxiv.org/pdf/1909.05478.pdf
- https://arxiv.org/pdf/1908.07162.pdf
- https://arxiv.org/pdf/1907.07590.pdf
- https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/
- https://aws.amazon.com/comprehend/features/
- https://www.OpenText.com/products-and-solutions/products/enterprise-content-management/ https://www.OpenText.com/products-and-solutions/products/enterprise-content-management/file- intelligence
- https://www.opentext.com/products-and-solutions/products/enterprise-content- management/documentum-platform
- https://www.OpenText.com/products-and-solutions/products/ai-and-analytics/
- https://www.OpenText.com/products-and-solutions/products/discovery
- https://www.adlibsoftware.com/adlibelevate.aspx
- https://www.adlibsoftware.com/products-and-services/offerings/file-analytics.aspx https://www.adlibsoftware.com/products-and-services/offerings/classification.aspx
- https://www.adlibsoftware.com/blog/2018/October/how-file-analytics-overcomes-legacy-challenges-of- document-classification.aspx
- https://www.adlibsoftware.com/products-and-services/integrations/microsoft.aspx
- https://www.recordpoint.com/classification-intelligence-records365/
- https://www.alfresco.com/ecm-software/document-management https://www.alfresco.com/ecm-software/search-and-insight-engine
- https://www.alfresco.com/ecm-software/alfresco-intelligence-services https://www.alfresco.com/ecm-software
- https://www.alfresco.com/information-governance
- https://cloud.google.com/natural-language/automl/docs/
- https://docs.microsoft.com/en-us/azure/machine-learning/service/overview-what-is-azure-ml https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning- initialize-model-classification
- https://www.ibm.com/uk-en/cloud/watson-natural-language-understanding https://www.ibm.com/watson/services/natural-language-classifier/

9. Legal Note