THE UNIVERSITY OF BRITISH COLUMBIA

School of Information
Faculty of Arts

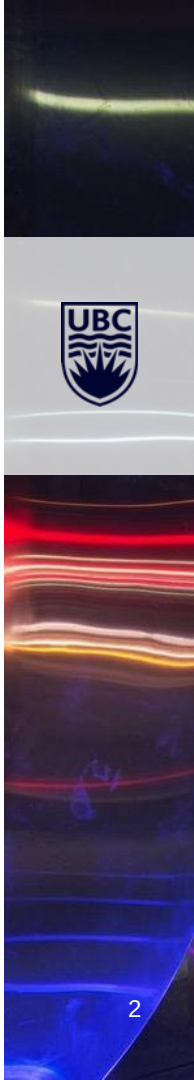
AN INTELLIGENT CLASS: THE DEVELOPMENT OF A NOVEL CONTEXT CAPTURING METHOD FOR THE FUNCTIONAL AUTO-CLASSIFICATION OF RECORDS

DECEMBER 15, 2023, NATHANIEL PAYNE

UBC iSchool

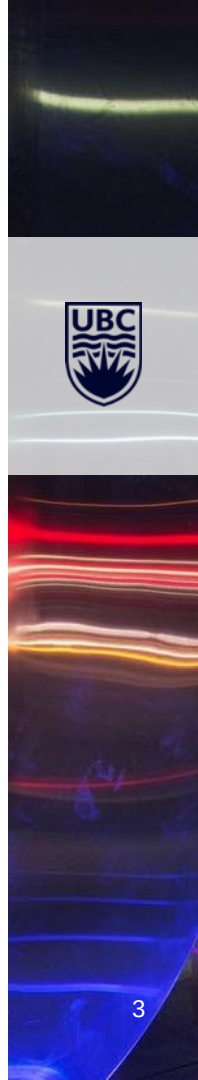
OUR JOURNEY

- Introduction & review of key issues
- Review of the research process & functional classification
- Review of the methodology & Enron data set
- Discussion of the theoretical framework (STACC)
- Review the output from classification & findings
- Conclusions, contributions & limitations



INTRODUCTION

- The need to classify records accurately is a core problem in many domains
- Historically, the functional classification of records was a manual process
- Once key attributes of a record were identified, record keepers classified a record according to its function (Administrative vs Operational)
- Due to significant growth in the volume of records, the classification of records **(assignment of a code to a record that associates with a functional class)** has become more complex and manually infeasible.



INTRODUCTION

- **OPEN RESEARCH QUESTION:** Can automated classification tools achieve categorization at the same level of accuracy and precision as a human classifier?

- **ARGUMENT:** The poor performance in the classification accuracy and precision of auto-classification exists because current tools are designed to read the content of a record rather than the archival diplomatic context of a record.

WHAT IS RECORDS CLASSIFICATION & WHAT ARE FUNCTIONS?

- Records classification is a simple process of assigning a code to a record

- Records classification is generally **primarily function-based**
    - Functional classification is used because activities are aggregations of functions and records are evidence of activities

- A **function**, is a set or series of activities (broadly speaking, a business process) which, when carried out according to a prescribed sequence, will result in an institution or individual producing the expected results in goods or services that it is mandated or delegated to provide.

WHY IS DETERMINING THE FUNCTIONAL CLASSIFICATION OF A RECORD IS HARD?

- Functional classifications are complex and differ between organizations

- It may be challenging to determine a single function by which a record should be classified, especially if there are many functions that are similar

- There may be uncertainty about whether an archivist has the skills or time to do so accurately

    - The first thing that a records manager learns is who is responsible for what. Without a deep understanding of an organization's functions, functional classification is hard

THE STEPS IN MY JOURNEY

1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude

PHASE 1 – SELECT MAIN MAIN DATASET & ORGANIZATION

- The main dataset and organization used for all research work was the **Enron email dataset**.

- Enron was the World's Leading Energy Company. Enron declared bankruptcy in December 2001.

- During the investigation into the bankruptcy, the original Enron email dataset, consisting of 92% of Enron's staff emails, i.e. 619,446 email messages in total, was posted to the web by the Federal Energy Regulatory Commission (FERC) in May of 2002.

PHASE 1 – FUNCTIONAL CLASSIFICATION SCHEME CREATION

- An **organizational systems analysis** was conducted to determine the key functions for Enron
- An functional classification scheme was created for the Enron records set
- The group of experts validated the functional classification scheme after multiple rounds of review
- Detailed scoping notes were created for each functional class which were reviewed by the human reviewers

CREATE SCOPE NOTES: EXAMPLE OF SCOPE NOTES FOR CLASS0028

Example Scope Note for Class 0028 Communications, Website Management

Records relating to developing and implementing corporate web site plans and strategies, coordinating web site content, and ensuring the relevance, accuracy, and timeliness of web site information. This primary also covers information presented on simple and dormant web sites not covered by an Operational Records Classification System (ORCS). Dormant web sites include abandoned web sites, web sites of defunct programs, and web sites of completed projects. Records types include correspondence; HyperText markup Language (HTML), Standardized General Markup Language (SGML), and Extensible Markup Language (XML) electronic documents and paper print outs of these documents; electronic forms; research materials; site designs; briefing notes; press releases; implementation plans; reports; and other types of records as indicated under relevant secondaries.

THE STEPS IN MY JOURNEY

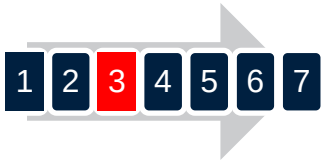
1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude

OUTPUT SAMPLE: ADMINISTRATIVE FUNCTIONAL CLASSES

0010Administration, General	0330Banks & Banking
0020Communications, General	0340Budgeting
0025Communications, Media Relations & PR	0350Expenditure Control, General
0028Communications, Website Management	0355Financial Reporting, Forecasting & Analysis
0030Disaster of Emergency Response Planning	0360Procurement & Contract Management
0040Executive Services, General	0365Revenue Control
0050Forms Management	0370Tax Reporting
0060Information & Privacy, General	0380Travel Authorization
0070Legal Administration	0390Treasury Management
0080Planning, Performance & Projects	0410Human Resource Management, General
0090Project Management	0420Benefits Administration
0095Records Management	0430Employee Supervision & Development
0096Reporting & Statistical Analysis	0440Leave & Time Reporting
0097Visit & Trip Planning	0450Occupational Safety, Health & Accidents
0110Equipment & Supplies, General	0460Staffing, Recruitment & Competitions

THE STEPS IN MY JOURNEY

1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude



PHASE 2 – HUMAN CLASSIFICATION

- Phase 2 required the completion of human classification and analysis work.

- Each expert was asked to manually classify 214 records from the Enron dataset into one of 80 functional categories agreed to in phase 1.

- It is important to note that the human classifiers had a hard task. This was because they were not familiar with the organizational system of Enron, or the main functions. Weeks of knowledge could not have filled the gaps. Thus, they started at a disadvantage.

FINDINGS FROM MANUAL CLASSIFICATION

In this table, *correct* refers to the case where two or more experts correctly and blindly classified a record to the category determined to match the generally accepted functional class for a record with similar features.

IMPORTANT: All expert records managers selected and agreed on the correct classification for a record only 7% of the time (n = 214).

Task Type	Incorrect	Correct	Total
Administrative	36%	38%	74%
Operational	19%	7%	26%
	55%	45%	100%

FINDINGS FROM MANUAL CLASSIFICATION

- There was a statistically significant difference between the accuracy demonstrated on administrative versus operational tasks for human classifiers

- Given that a record was an operational record, the probability that a manual classifier would make an error during the classification task was 75%.

- For administrative tasks, given that a task was administrative, the probability of making a classification error for this task was only 48%.

- Such findings reinforce the need for records managers to understand an organization's operational context before beginning the operational classification process. Note that they did have a disadvantage of their knowledge of the system, which was a limitation.

FINDINGS & DISCUSSION FROM MANUAL CLASSIFICATION

- There is a critical need for good-quality scope notes during the functional classification process
- Gaps between the administrative and operational understanding can impair the records classification process.
- Small buckets create big problems. Human classifiers had a hard time classifying records as the number of buckets increased
- Human reviewers have a hard time determining whether the functional classification is exhaustive.

THE STEPS IN MY JOURNEY

1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude

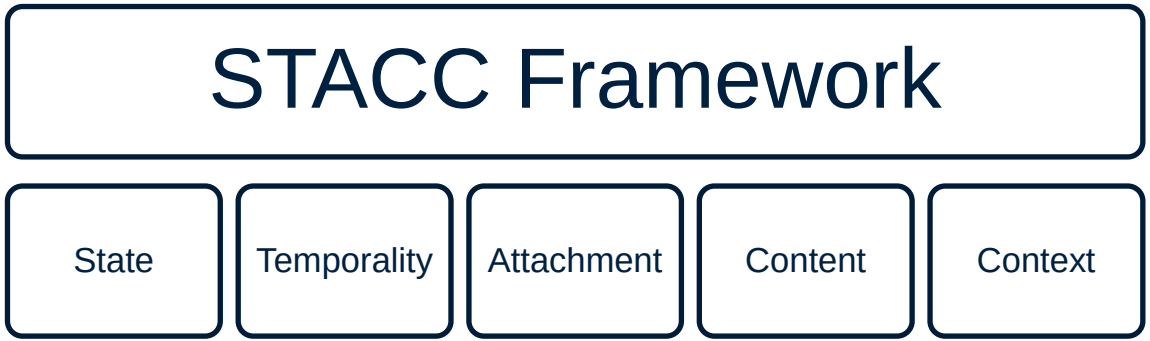
UBC iSchool

UTILIZING CONTEXT

- The definition of context that informs this thesis is that from archival diplomatics.
- At the heart of diplomatics lies the idea that all records can be analyzed, understood, and evaluated in terms of a system of formal elements that are universal in their application and decontextualized in nature (Duranti, 1996, 1996, 1997, 1998, 1998a, 1999)
- Electronic records, just like every traditional record, have the following key elements: Persons, Acts, Archival Bond, Content, Context, Form, Medium (Duranti, 1999).
- My goal, was to create a process that extracted elements from the record that used the 7 key elements of a record, and used these elements to classify the record to a function

CONTEXT & THE STACC FRAMEWORK

- I have proposed a new framework that aims to include all aspects of traditional archival diplomatics' concept of a record, including the persons, acts, archival bond, content, context, form, and medium.
- This framework, is an **initial attempt** at combining computational components with archival diplomatics components to enable robust analysis and auto-classification.
- This is a beginning - a starting point - for both creating a classification system and for classifying according to a computational framework



THE 5 ELEMENTS OF THE STACC FRAMEWORK

- The **state** of a record can be defined as the sum of its formal attributes, including its rules of representation, and which capture its physical, informational form, and medium.

- **Temporality** can be defined as the organizational, functional, and operational circumstances surrounding a record's creation, receipt, storage, or use and its relationship to other materials.
    - Temporality is **uniquely related to time**

THE 5 ELEMENTS OF THE STACC FRAMEWORK

- ***Attachment*** can be defined as the theoretical relationship each record has with the other records produced as part of the same transaction or activity and located within the same grouping.

- ***Content***, including the keywords that are present within the subject and text of records, represents a connection between scope notes and transactional items.

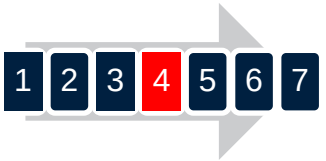
- **Context** can be defined as the organizational, functional, and operational circumstances surrounding a record's creation, receipt, storage, or use.

MAPPING ARCHIVAL DIPLOMATICS -> STACC

Archival Diplomatic Contextual Elements	What To Be Operationalized?	STACC Framework Components For Computability
Juridical-Administrative	Juridical Category	Public/Private (Internal vs External)
Provenancial	Record Persons	From (Author), To (Addressee), Writer (Author)
	Place	Internal / External (Binary)
	Known Source	Yes / No (Binary)
	Transmission Status (Primitive)	Yes / No (Binary)
Procedural	Simple vs Collegial Act	Binary (0/1); Categorical
	Transaction vs Narrative / Supporting Document	Categorical (Transaction vs Narrative / Supporting)
Documentary	Index	Unique Record ID
	Temporality (Time)	Date, Record Timestamp (UNIX), Day (Categorical), Month (Categorical)
	Fond	Email Category (Folder)
	Subject Classification	Subject Keyword (Theme)
	Archival Bond	Network Score (Eigenvalue Centrality)
Technological	Transmission Status (Completeness & Effective)	Binary (1/0)

UBC iSchool

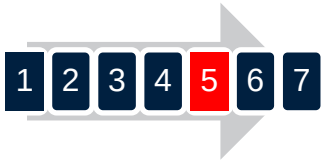
THE STACC FRAMEWORK: LINKING COMPUTATIONAL SCIENCE AND ARCHIVAL DIPLOMATICS



Archival Diplomatics Components of A Record	Theoretical Metadata	STACC Framework Components For Computability
Form (the rules of representation that allow for the communication of the message)	Chained (Y/N), Completed (Delivered = Y/N), Authenticity (Validated/Not)	State (Form, Medium, Memory Preservation)
Medium (the physical carrier of the message)	Message Type (Email)	State (Form, Medium, Memory Preservation)
Documentary Context (the documentary framework in which the action takes place)	Record Timestamp (Hour), Public/Private (Internal vs External), Day (Categorical), Month (Categorical)	Temporality (Time & Connectivity);
Persons (the entities acting by means of the record)	From (Author), To (Addressee), Writer (Author)	Attachment (Connectivity, Centrality)
Archival Bond (the relationship that links each record to the previous and subsequent one and to all those which participate in the same activity)	Network Score (Eigenvalue Centrality), Email Category (Folder Name),	Attachment (Connectivity, Centrality)
Acts (the exercise of will that originates the record as a means of creating, maintaining, changing, or extinguishing situations)	Transaction derived from functional thesauri (an action between two or more persons, aiming to change the relationship existing between them) vs Narrative / Supporting function), Simple/Collegial (Act)	Context (Circumstance, Acts)
Content (the message that the record is intended to convey)	Keywords (From Scope Notes), Subject Keyword (Single)	Content (Variety, Uncertainty)

THE STEPS IN MY JOURNEY

1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude



PHASE 3 – AUTO CLASSIFICATION

- After completing phase two, the elements from STACC were extracted from each record and the data was cleaned.
- The auto-classification algorithm that was developed was run against all distinct and cleaned records in the set (final 200,399 records).
- After the model was trained, the model version with the highest precision score on the training data was run against the 214 records in the test set post-training and optimization (to compare the human vs machine)
- The auto-classification algorithm leveraged the random forest (RF) classifier, using 80/20 training a testing with one thousand iterations.

WHAT IS AUTO-CLASSIFICATION

- Auto-classification of records refers to the process of using computational methods to automatically assign a classification code to a digital record

- The high cost of manual classification has encouraged the development of new auto-classification methods (Golub, 2006; Santangelo, 2009).

- Functional auto-classification of records refers to using computational methods to automatically organize records based on their activity or program (function) and place them into such an organization (Payne, 2022; Payne & Baron, 2017).

HOW DO WE COMPLETE AN AUTO-CLASSIFICATION TASK

- To complete an auto-classification task, a computational system must be able to read records as they are made or received, categorize those records by distinct attributes, and assign a record to a functional class automatically.

- To auto-classify a record, a computational device must utilize detailed information from the record to assign the record to the appropriate function.

- The computer must use many meta-data elements to make the assignment.

FINDINGS FROM AUTO-CLASSIFICATION

- The auto-classification process accuracy exceeded that of the manual classification process, with an accuracy of 63% for the records sample obtained during the validation step (n = 214).
- This is compared to a median accuracy of 49.3% for the overall group of reviewers. The best manual classifier, for reference, had an accuracy score of 55.6%.
- There was also a statistically significant difference between the accuracy demonstrated on administrative versus operational tasks

Task Type	Incorrect	Correct	Total
Administrative	26%	49%	74%
Operational	12%	14%	26%
	37%	63%	100%

FINDINGS & DISCUSSION FROM AUTO-CLASSIFICATION

- Computers need help with unstructured data, including the acquisition of it and utilization.
- Extracting features from records is challenging and time consuming hard
- Intuition was relied upon for decision-making by human reviewers. Processing intuitive context is challenging for a computer.
- A larger number of buckets slows processing time, especially if those buckets have sparse information or few examples.
- Records with information that needed clarification or did not have identifiable features were challenging to classify.

THE STEPS IN MY JOURNEY

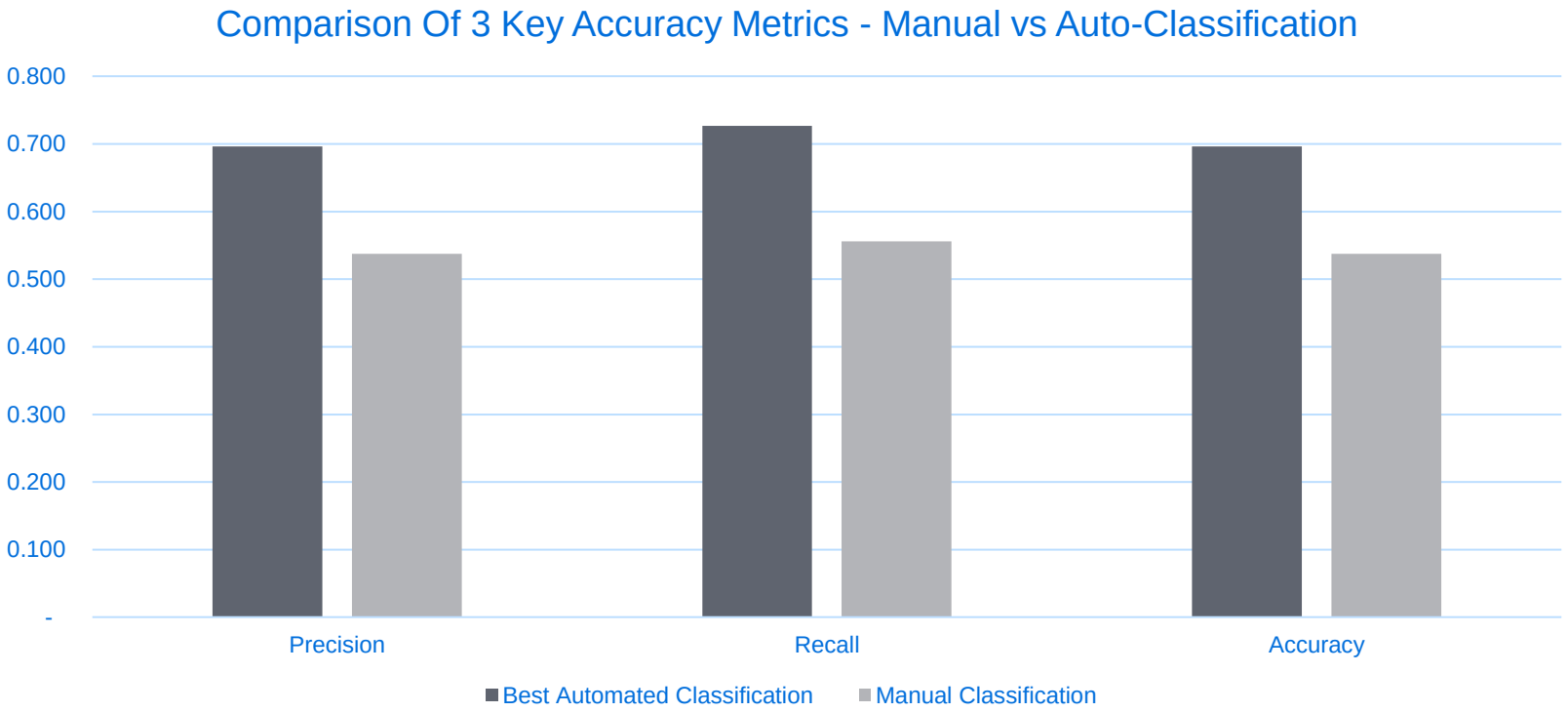
1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude

COMPARING MANUAL & AUTO-CLASSIFICATION

- The auto-classification model was tuned through successive iterations, with weight optimizations, as well as other explanatory variables.
- This included tuning the LDA model and relevant keywords for operational records, while removing redundant keywords manually that provided misleading results.
- Key variables: The **persons**, **keywords** for the subject and body that maps to functions, **private vs public**, and eigenvalue centrality (used as a proxy for the archival bond), where the key variables

		Precision	Recall	Accuracy
Analysis Method	Manual Classification	0.49	0.43	0.49
	Auto-Classification	0.63	0.55	0.63
	Auto-Classification Final Tuned	0.70	0.60	0.70

COMPARISON OF 3 KEY ACCURACY METRICS



UBC iSchool

THE STEPS IN MY JOURNEY

1. Create a functional classification that matches the Enron environment
2. Test this functional classification with classification experts
3. Let experts manually classify 214 records based on the functional classification
4. Create a schema that maps archival diplomatics to computer science
5. Conduct an auto-classification process using STACC
6. Compare the results from the manual and auto-classification
7. Complete a meta-analysis with the expert reviewers & conclude

FINDINGS FROM META-ANALYSIS OF REVIEWERS

- When organizational knowledge (knowledge of an organization's functions) is weak, the ability to classify empirically becomes more difficult.
    - This difficulty occurs because human classifiers rely intrinsically on organizational knowledge to improve the accuracy and precision of their classification decisions.
- Using the elements from archival diplomatics, primarily when these elements are used to model the scope notes, was helpful. Scope notes were key links between the content and the context for human classifiers.
- In situations where a function was closer to other functions, especially in the operational functions, the ability for humans to classify records decreased.

FINDINGS FROM META-ANALYSIS OF REVIEWERS

- From a classification accuracy perspective, the content of the message was found to be necessary, but not for the content of what the words communicated.
    - As noted by the group of experts, as they reviewed the record's content, they were inherently and ultimately reading for archival diplomatic context, not aboutness.
- For the manual reviewer, the process of classification is memory intensive. If the reviewer does not have the context against the records set, their ability to determine accurately the functional class is impaired.
- As the number of rules increases, the human classifier is at a distinct disadvantage because there are more things to track.

SIGNIFICANCE OF THE STUDY

- This is the first publicly available research that created a functional classification scheme for the Enron dataset, and the first to validate such a scheme with a group of experts.

- This research is the first attempt **to operationalize the definition of the context of a record that leverages an archival diplomatics' definition of context**.
    - It is a starting point that will be built upon in the future for research relating to context.

- This work is the first to use a mimicry methodology to develop a method and computational framework within Archival Science that mimics a human classifier.

CONCLUSIONS

- The auto-classification task was able to classify more accurately than a human manual classification classifier team by nearly 17% on average ($p < 0.0001$) while improving the classification accuracy for administrative and operational records.

- There is a significant variance between records managers during the manual classification process, with statistically significant differences in their ability to accurately classify administrative and operational records.

- This thesis rejected the null hypothesis that using a record's context will not improve the accuracy, precision, and recall of auto-classification methods compared to the accuracy, precision, and recall attainable by human classifiers.

LIMITATIONS

- While the research uses a group of experts, the study only focuses on one significant, publicly available record set. This choice has been made because of the limited availability of global records datasets.

- The other limitation relates to using human subjects to validate the functional classification scheme manually. The human classifiers had a distinct disadvantage as their starting point when creating the functional classification.

    - A different set of reviews could produce different classification results. It is arguable that the results would be directional to what was observed here. That said, this is a first step.

LIMITATIONS

- While the machine classifier did outperform the human classifier, it is not clear and not possible at this point to determine how much the improvements in the classifier were due to the additional archival diplomatic context and how much of the gain in accuracy was due to the superior information processing power of the machines.

- Future research needs to complete, and is intended to complete, an exhaustive review of the optimal settings for machine learning algorithms' testing on a large dataset. It is possible that other machine learning methods could yield better results

    - My research goal is not to complete a simulation of all possible feature extraction settings for each record's context elements

THE UNIVERSITY OF BRITISH COLUMBIA

School of Information
Faculty of Arts

THANK YOU & QUESTIONS

DECEMBER 15, 2023, NATHANIEL PAYNE

REFERENCES

Duranti, L. (1995) Reliability and Authenticity: The Concepts and Their Implications. Archivaria, 1 (39) (1995): 5– 10

Duranti, L. (1996) 'Archives as a place', Archives and Manuscripts, vol. 24, no, 2, pp. 241- 255.

Duranti, L. (1997) "The Archival Bond," Archives & Museum Informatics 11, nos. 3-4: 213-18

Duranti, L. (1998). Diplomatics: New Uses for an Old Science. Lanham, MD, and London: The Scarecrow Press. Archivaria 28. Part 1.

Duranti, L. (1998a).Diplomatics: New Uses for an Old Science. Lanham, MD, and London: The Scarecrow Press, Archivaria 30. Part 3

Duranti, L. (1998c). Diplomatics: New Uses for an Old Science. Lanham, MD, and London: The Scarecrow Press, Archivaria 6-32. Part 5

Golub, K. (2006). "Automated subject classification of textual web documents." Journal of Documentation. 62(3):350-71.

Santangelo, J. (2009). Rise Of The Machines. Information Management, 43(6), 22.

Payne, N. (2022, December). An Intelligent Class: The Development Of A Novel Context Capturing Method For The Functional Auto Classification Of Records. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 2546–2555). IEEE.

Payne, N., & Baron, J. R. (2017, December). Auto-categorization methods for digital archives. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2288–2298). IEEE.