Journal of Library and Information Studies 23:2 (December 2025) pp.119-143
https://doi.org/10.6182/jlis.202512_23(2).119

Exploring Class Mapping as Data Fusion Technique in Machine Learning for Research Classification

Chien-Chih Huang[1], Kuang-Hua Chen[2]

Abstract

Access to sufficient, high-quality data is essential for effectively training and validating machine learning classifiers. This study investigates class mapping as a data fusion strategy to enhance training data for research classification. Two versions of the Australian and New Zealand Standard Research Classification, ANZSRC 2008 FoR and ANZSRC 2020 FoR, are used to organize 179,431 documents from eight institutional repositories into plain and mapped datasets. Each dataset is divided into subsets corresponding to the division, group, and field levels of the classification schemes. Results show that 49% to 63% of documents are successfully mapped between schemes. Classifiers by Support Vector Machines (SVM), SciBERT, ModernBERT-base, and ModernBERT-large are trained to assess the effectiveness of this data fusion approach on classification performance. All models show improved performance at the three levels. ModernBERT-large achieved the greatest performance gains, with the improvements in validation F1 scores of 1.0% and 2.5% at the division level, 4.4% and 2.2% at the group level, and 9.9% and 11.5% at the field level. An emergent ability was observed, as performance in non-augmented classes improved with ModernBERT-large but not with ModernBERT-base. Overall, this study demonstrates that class mapping effectively enriches training datasets, enhances classification performance, and underscores the importance of model size and architecture. These findings offer a practical and scalable strategy for improving machine learning performance in research classification tasks.

Keywords: Interoperability; Inter-concept Mapping; Machine Learning; Research Classification

1. Introduction

Research classification is served for bibliographic and administrative purposes (Hjørland & Gnoli, 2022) for various or specific regions or disciplines. While evaluating or comparing the research output collections from incompatible schemes, arranging collections into a unified classification scheme is an ordinary approach. The correspondence table is a common method in Library and Information Science (LIS) to interoperate two schemes. We will use the correspondence table as a data fusion strategy to develop classifiers for two research schemes. The Australian and New Zealand Standard Research Classification (ANZSRC) is a classification system developed to measure and analyze the research and experimental development (R&D) statistics in Australia and New Zealand (Australian Bureau of Statistics, 2020b; Commonwealth of Australia and New Zealand, 2020). ANZSRC was first released in 2008 and revised in 2020 to keep up with the pace of contemporary research. Fields of Research (FoR) is one of three classifications

[1,2] Department of Library and Information Science, National Taiwan University, Taipei, Taiwan
* Corresponding Author: Kuang-Hua Chen, E-mail: khchen@ntu.edu.tw

in the ANZSRC and the fields are categorized according to "common knowledge domains and/or methodologies" (Australian Bureau of Statistics, 2020b). ANZSRC FoR is not only used in Australia and New Zealand but also globally employed by Springer Nature SciGraph (Pasin, 2017) and Dimensions (Digital Science and Research Solutions, 2022a, 2022b). Dimensions stated that ANZSRC FoR encompasses all academic disciplines at a general level, allowing for comparisons across various research areas. Although ANZSRC FoR is designed to include all research areas, the scheme is inevitably outdated and is revised to keep the immediacy. However, a forthcoming document classified in the revised scheme cannot be directly compared with those in the original scheme. Only 9% of the documents in our dataset are classified in both two revisions. Collection evaluation necessitates reclassification between ANZSRC 2008 FoR (FoR2008) and ANZSRC 2020 FoR (FoR2020), where reclassification efforts can be optimized through systematic class mapping.

This study aims to explore the interaction among three interconnected components: bibliographic records, class mapping, and classification models. The primary objective is to enrich plain datasets through class mapping between schemes. The second one is to systematically evaluate various machine learning algorithms to identify optimal classifiers for both plain and mapped datasets. The third one is to examine the factors that influence improvements in classification performance. The class mapping established in the correspondence table delineates relationships between classes in one scheme and their counterparts in the other scheme, with each class potentially corresponding to multiple classes across schemes. The FoR2020 scheme is so updated that the records of some classes are insufficient for training classifiers. This shortage is partially mitigated by augmenting records from FoR2008. The field of AI has repeatedly "reinvented the wheel" to address challenges that the LIS field had already developed solutions for years earlier (Dahlberg, 1993). Interoperability, a concept implemented in LIS well before the advent of the internet (Zeng, 2019), is operationalized in this study through a correspondence table to enrich bibliographic records for developing machine learning classifiers.

2. Related Studies

The first section introduces the research classification and the ANZSRC. The second section presents the text classification by machine learning. The document and class interoperability are discussed in the third section.

2.1 Australian and New Zealand Standard Research Classification

Research classification systems are purposed for reporting research activities (Hjørland & Gnoli, 2022) by various organizations or countries. Some classification systems (e.g., The Flemish Research Discipline Classification, Vlaamse Onderzoeksdiscipline Standaard, VODS) may be termed "discipline classification" but essentially refer to research classification. Disciplines are tightly connected to the phenomenon of teaching and research, which are two main missions of scholars in modern research universities (Hammarfelt, 2020). As to the teaching mission,

the International Standard Classification of Education (ISCED) is one of the universal education classification systems organized by the education levels and fields (United Nations Educational, Scientific and Cultural Organization, 2015), which differs from the research activities. Hider and Coe (2022) mapped university faculty structure to the bibliographic, education, and research classification systems. 56.7% of the university structures are mapped to the Dewey Decimal Classification (DDC), 49.8% to the Library of Congress Classification (LCC), 61.2% to the Australian Standard Classification of Education (ASCED), and 54.2% to ANZSRC. The varied ratios depict that those classification systems are not aligned since the disciplines and fields are "sliced and diced" in the universities. Mapping collections across institutions, libraries, or disciplines requires a unified system. Research classification offers a viable solution, as it strikes a balance between the specificity of bibliographic classification and the broad scope of educational classification.

This study adopts the ANZSRC FoR as the targeted research classification system, as it aligns with the Frascati Manual of the Organisation for Economic Co-operation and Development (OECD) (Australian Bureau of Statistics, 2020b; Hjørland & Gnoli, 2022; OECD, 2015) and could be crosswalked to other classification schemes. The Frascati Manual of the OECD defines the most globally recognized standards and recommendations to collect and report comparable statistics about research and experimental development. The ANZSRC is internationally applicable and leveraged or referred by other schemes such as Canadian Research and Development Classification (CRDC) in Canada (Legendere, 2019), or the Flemish Research Discipline Classification Standard in Belgium (Vancauwenbergh & Poelmans, 2019). Legendere (2019) asserts that the reference to the Frascati Manual and ANZSRC aims to increase computability, collaboration, and international standards alignment. The correspondence table enables ANZSRC to be interoperable with classification schemes based on the Frascati Manual, providing a foundation for mapping to additional schemes. The databases such as Springer Nature SciGraph (Pasin, 2017) and Dimensions (Digital Science and Research Solutions, 2022a, 2022b) employ ANZSRC FoR to classify the curated documents. Bornmann (2018) manually inspected the classification results of his 199 articles in Dimensions, and he found that "most of the papers seem misclassified." (p. 639) However, automatic classification studies for ANZSRC FoR are needed since manual classification for a huge number of documents is infeasible. Our study increases the number of training documents to improve the classification performance.

ANZSRC was established jointly by the Australian Bureau of Statistics and Statistics New Zealand, and contains three classifications for the measurement and analysis of research and experimental development in Australia and New Zealand. Preceded by the Australian Standard Research Classification (ASRC) of 1998, the ANZSRC was published in 2008 and revised in 2020. Three ANZSRC classifications are: (1) Type of Activity (TOA), which categorizes the types of research effort; (2) Fields of Research (FoR), which categorizes the common knowledge domains and/or methodologies; (3) Socio-economic Objective (SEO), which categorized

the intended purpose or outcome perceived by the authors (Australian Research Council, n.d.). ANZSRC FoR has a hierarchical structure of three levels, which are named Divisions, Groups, and Fields. FoR2020 establishes the Indigenous Studies Division, eliminates the Technology Division, and tears the Medical and Health Science into two more focused Divisions (Commonwealth of Australia and New Zealand, 2020). A division class is represented by a two-digit number, and a group class is represented by a four-digit number, of which the first two digits stand for the belonging division class. A field class is represented by a six-digit number in which the first two digits refer to the belonging division class, and the first four digits signify the belonging group class. The division classes of FoR2008 are numerated from 01 to 22. In contrast, the division classes of FoR2020 are numerated from 30 to 52. Class number ending in 99 denotes a miscellaneous class, which is designed to include not-elsewhere-classified topics, including cutting-edge discovery. Macauley et al. (2011) discovered that disproportionately high numbers of theses in some group classes are classified into miscellaneous classes, and they suggested the classification by authors, as well as the update of the legacy scheme. FoR2008 has 22 divisions, 157 groups, and 1,238 fields. FoR2020 has 23 divisions, 213 groups, and 1,967 fields. The number of groups or fields growing over time indicates the later schemes extended to include new topics. The Australian Bureau of Statistics offers correspondence tables including the table between FoR2008 and FoR2020, from which the mapping relations between schemes are derived. The Research Excellence Branch of the Australian Research Council conducted a manual classification task (Macauley et al., 2011) for classifying 9,051 Ph.D. thesis into FoR2008. 47.6% of the theses were allocated with 2 labels, 26.8% had 1 label, and 25.6% had 3 labels. They suggested that the codes should be assigned by the authors, who are familiar with the text content and can allocate accurate labels. Authors, however, may not be familiar with the scheme and may classify better by suggesting plausible classes.

2.2 Automatic text classification

The advancement in AI technology is capable of assisting the classification tasks for knowledge organization systems (KOS), and the natural language processing (NLP) methods are overtly amended as of the 2010s (Collobert et al., 2011). The "representation" is one of the changing features of modern NLP. Each linguistic entity, such as a word (e.g., Mikolov, Chen, et al., 2013), a phrase (e.g., Mikolov, Sutskever, et al., 2013), as well as a sentence or document (e.g., Le & Mikolov, 2014), is represented by a real-valued vector of the distributed representation, which is contrary to the distributional representation such as term frequency or TF-IDF. Word2vec and fastText (Bojanowski et al., 2017) are popular static embedding models that can deal with various semantic tasks, such as text classification, but cannot tackle polysemy. By leveraging ELMo (Peters et al., 2018) and Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2019) generates dynamic embedding in which the vectors of the entities are subject to the context words. A polysemic word is represented by relatively dissimilar vectors if the word performs distinctive semantics in different contexts. Modern NLP

models are capable of distinguishing not only the polysemy but also the sequential order of words in a sentence that traditional bag-of-words models cannot. Garcia-Silva and Gomez-Perez (2021) built multi-label FoR2008 classifiers with BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), GPT-2 (Radford et al., 2019), SVM (Sebastiani, 2002), and fastText (Bojanowski et al., 2017) on the SciGraph dataset. The SciBERT model achieved the highest F1 score at the division level. Among nine selected group-level classes, the top-performing classifiers varied: SciBERT outperformed in five groups (Biological, Medical and Health, Chemical, Mathematical, and Computer Sciences), native BERT excelled in two groups (Language and History), and SVM led in two groups (Built Environment and Creative Arts). ModernBERT (Warner et al., 2024), a BERT variant like SciBERT, features architectural enhancements and is pretrained on scientific literature and web data, making it well-suited for our research classification tasks. ModernBERT provides a large model variant, which theoretically offers improved performance. This enables a clearer assessment of classifier effectiveness when trained on both plain and mapped datasets. Wu et al. (2021) trained traditional machine learning models, including Multinomial Logistic Regression (MLR), Multinomial Naive Bayes (MNB), K-Nearest Neighbors (KNN), and SVM on records from Research Data Australia (RDA) at the division level of the FoR2008. They suggested that the group or field level classifiers are more suitable for practical use. S. Zhang et al. (2023) employed ChatGPT on the RDA dataset and demonstrated that ChatGPT did not generally outperform MLR or KNN models. They prompted only the division classes and exemplar articles since prompting all field-level class headings is too lengthy for ChatGPT 3.5. ChatGPT often generated hallucinations that documents are classified into non-existent classes in our trial. The above automatic text classification studies regarding the ANZSRC FoR dealt with division-level or group-level classification, but no studies achieved field-level classification of more than 1,000 classes. Our study accomplished the classifiers of three levels with the above-mentioned and additional large language models (LLM). Arhiliuc et al. (2025) conducted a multi-label classification of journal article abstracts from the Web of Science into 42 OECD FORD classes using BERT, SVM, SPECTER, and GPT-3.5. They found that BERT outperformed the other models, followed by SVM+TFIDF, SVM+SPECTER, and GPT-3.5. They highlighted the "scarcity of labeled multidiscipline data", and address the scarcity by aggregating the records from multiple repositories and mapping classifications from other schemes.

2.3 Interoperability

A scheme is revised to reflect advancements in science, creating the need to reclassify labeled articles from the legacy scheme into the new one. Porter et al. (2023) combined bibliometric clustering, venue subject, manual class assignment, and direct mapping at the field-level classes from FoR2008 to FoR2020 in Dimensions, which employs SVM to classify at the division and group level of the schemes. However, they do not list any classification performance metrics. L. Zhang et al. (2022) examined the article-level classification consistency among three databases: (1) Web of Science subject categories (WoS SC);

(2) Dimensions FoR classes, which is derived from FoR2008; (3) subject classification of Springer Nature (SNSC). WoS SC is generated from the citation relation. SNSC is labeled by the authors. Articles are mapped or classified into OECD FOS (Field of Science and Technology) 2007, which is composed of 6 major categories and 43 minor categories. The results showed that single-category assignment in WoS SC is generally inappropriate, which confirmed the viewpoint of Macauley et al. (2011) that multi-classes are more appropriate to describe a document. Their study demonstrated the re-classification process via mapping relation, which defined the relations between two documents as identical, partially identical, and inconsistent. Our study would formalize the relations. The classification in the three databases is greatly inconsistent in that only 27% of papers had identical fields between the machine-generated Dimensions FoR code and human-judged SNSC. Since the articles authored by Bornmann (2018) in Dimensions are mostly misclassified, it would be ideal to boost the classification performance before the examination of the consistency by class mapping. In contrast, the classification performance may be improved by class mapping as our study would demonstrate.

3. Class Mapping Relation

Australian Bureau of Statistics (2020b) publishes "ANZSRC 2020 correspondence to ANZSRC 2008" (Australian Bureau of Statistics, 2020a) that enumerates the class mapping between ANZSRC 2020 FoR and ANZSRC 2008 FoR. A correspondence relation is dyadic, meaning that a class in one scheme corresponds to a single class in the other scheme. Given the correspondence relation is denoted as $\leftrightarrow$, the correspondence relation is denoted as $cls_i \leftrightarrow cls_j$, which means that class i maps to class j. $\leftrightarrow$ is symmetric that $cls_i \leftrightarrow cls_j$ is equivlant to $cls_j \leftrightarrow cls_i$. The class set of the ANZSRC 2008 FoR is denoted as $CS^{FoR2008} = \{cls_i | \forall cls_i \in FoR2008\}$. Similarly, $CS^{FoR2020} = \{cls_j | \forall cls_j \in FoR2020\}$. The mapping matrix $MAP \in \mathbb{Z}_2^{|CS^{FoR2008}| \times |CS^{FoR2020}|}$ is defined as:

$$MAP_{i,j} = \begin{cases} 1 & cls_i \leftrightarrow cls_j \text{ is in the correspondnece table} \\ 0 & cls_i \leftrightarrow cls_j \text{ is not in the correspondnece table} \end{cases}$$

The mapping relation between classes of two schemes is derived and categorized with the mapping matrix. The row sum of the mapping matrix MAP with respect to cls_i, i.e., $rowsum(MAP, i) = \Sigma_{k=1}^{|CS^{FoR2020}|} MAP_{i,k}$, is the number of $FoR2020$ classes to which cls_i is mapped. The column sum with respect to cls_j, i.e., $colsum(MAP, j) = \Sigma_{j=1}^{|CS^{FoR2008}|} MAP_{k,j}$, is the number of $FoR2008$ classes to which cls_j is mapped. The relation of a class pair is identified by three conditioned variables: (1) $MAP_{i,j}$, (2) $rowsum(MAP, i)$, and (3) $colsum(MAP, j)$. The previous studies had identified four kinds of class relation: (1) equivalence, (2) inclusion, (3) is about, and (4) union (Dahlberg, 1998; Meo-Evoli et al., 1998). We rename "is about" to "overlay" for readability. Since the union relation can be fully expressed by inclusion relation, union relation is omitted in Table 1. In addition, the disjoint relation is appended to describe the non-mapped type.

The above four kinds of relations are further simplified into three types of relations by jointly considering the mapping direction as shown in Table 2. Three types of class relation are (1) non-mapped, (2) possibly-mapped, and (3) definitely-mapped. Our study applies the definitely-mapped relation to propagating documents' classification

Table 1. Four Kinds of Class Relation

Kind	Denotation	Conditioned variables		
		$MAP_{i,j}$	$rowsum(MAP, i)$	$colsum(MAP, j)$
Disjoint	$Cls_i \parallel Cls_j$	0	(ANY)	(ANY)
Equivalence	$Cls_i = Cls_j$	1	1	1
Overlap	$Cls_i \otimes Cls_j$	1	>1	>1
Inclusion	$Cls_i \subset Cls_j$	1	1	>1
Inclusion	$Cls_i \supset Cls_j$	1	>1	1

Note. $Cls_i \in CS^{AFoR2008}$, $Cls_j \in CS^{FoR2020}$

Table 2. Three Types of Class Relation

Type	Mapping direction	
	FoR2008 to FoR2020	FoR2020 to FoR2008
non-mapped	$Cls_i \parallel Cls_j$	$Cls_i \parallel Cls_j$
possibly-mapped	$(Cls_i \otimes Cls_j)$ OR $(Cls_i \supset Cls_j)$	$(Cls_i \otimes Cls_j)$ OR $(Cls_i \subset Cls_j)$
definitely-mapped	$(Cls_i = Cls_j)$ OR $(Cls_i \subset Cls_j)$	$(Cls_i = Cls_j)$ OR $(Cls_i \supset Cls_j)$

Note. $Cls_i \in CS^{FoR2008}$, $Cls_j \in CS^{FoR2020}$

labels from one scheme to the other one. 74% (802/1,081) of FoR2008 non-miscellaneous field-level classes are definitely mapped to FoR2020. On the contrary, 85% (1,499/1,754) of FoR2020 non-miscellaneous field-level classes are definitely mapped to FoR2008. The definitely-mapped relation is the only type employed to augment the datasets. Porter et al. (2023) discovered that 80% of FoR2008 codes could be mapped directly to codes of FoR2020. Our study provides a more accurate ratio with the operation of the mapping matrix.

4. Research Questions

(1) What quantitative increases are observed in the number of classes and documents when transitioning from plain datasets to mapped datasets?

(2) What is the comparative performance of traditional machine learning and deep learning algorithms across the three levels of the ANZSRC Fields of Research (FoR) classification scheme?

(3) How does class mapping influence classification performance metrics?

5. Method

The bibliographic records containing the metadata fields of title, abstract, as well as ANZSRC 2008 FoR or ANZSRC 2020 FoR, are harvested from eight repositories. The records are organized and/or mapped through the correspondence table into four datasets. The classification algorithms are trialed and presented with the finest ones in the result section.

5.1 Dataset and mapping

The repositories in Open Access Australasia (https://oaaustralasia.org/directory-type/open-repositories/) were examined for the downloadability of bibliographical records via OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Bibliographical records were downloaded from eight institutional repositories, which are listed in Table 3, between July 27th, 2023, and July 31st, 2023. The documents were selected with the following criteria: (1) records with explicit codes and names of the FoR2008 and FoR2020 classes; (2) only the genre of the dissertation, thesis, conference, and journal article, as well as the proper description of an academic book; (3) more than 200 characters of the cleaned abstract text. The cleaning procedure removes the irrelated text, such as DOI, funding, acknowledgment, copyright announcement, and embargo period.

A document from the plain dataset of one scheme is incorporated into the mapped dataset of the other scheme if and only if all classes of that document are definitely mapped. The following is the mapping procedure for a document. The ground truth of a document's classes is represented by $truth^{FoR2008} \in \mathbb{Z}_2^{|CS^{FoR2008}|}$ for FoR2008, or $truth^{FoR2020} \in \mathbb{Z}_2^{|CS^{FoR2020}|}$ for FoR2020.

Table 3. Harvested Repositories

Institute	OAI-PMH URL	fetched records	FoR2008 records	FoR2020 records
Australian National University	https://openresearch-repository.anu.edu.au/oai/request	270,902	92,108	4,742
James Cook University	https://researchonline.jcu.edu.au/cgi/oai2	54,681	36,170	17,126
Lincoln University	https://researcharchive.lincoln.ac.nz/dspace-oai/request	7,432	3,881	848
Massey University	https://mro.massey.ac.nz/oai/request	15,691	1,022	1,641
University of Canterbury	https://ir.canterbury.ac.nz/oai/request	23,153	3,810	5,555
University of New England	https://rune.une.edu.au/uneprodoai/request	31,492	26,934	17,456
University of Southern Queensland	https://eprints.usq.edu.au/cgi/oai2	29,356	0	26,571
Victoria University	https://vuir.vu.edu.au/cgi/oai2	30,985	25,930	3,734

$$truth_k^{For2008} = \begin{cases} 1 & \text{if the document is labeled with the FoR2008 class } k \\ 0 & \text{otherwise} \end{cases}$$

$$truth_k^{For2020} = \begin{cases} 1 & \text{if the document is labeled with the FoR2020 class } k \\ 0 & \text{otherwise} \end{cases}$$

The mapped vector is $mapped^{FoR2020} \in \mathbb{Z}_2^{|CS^{FoR2020}|}$ or $mapped^{FoR2008} \in \mathbb{Z}_2^{|CS^{FoR2008}|}$. Let $\Psi^{FoR2020} = truth^{FoR2008} \times MAP$ or $\Psi^{FoR2008} = truth^{FoR2020} \times MAP^{Tranpose}$. If $\Sigma_{k=1}^{|CS^{FoR2020}|} \Psi_k^{FoR2020}$ equals to $\Sigma_{k=1}^{|CS^{FoR2008}|} truth_k^{FoR2008}$, it implies that all FoR2008 classes in that document are definitely-mapped. As a result,

$$mapped_j^{FoR2020} = \begin{cases} 1 & \text{if } (\Psi_j^{FoR2020} > 0) \text{ and} \\ & \left(\sum_{k=1}^{|CS^{FoR2020}|} \Psi_k^{FoR2020} = \sum_{k=1}^{|CS^{FoR2008}|} \text{truth}_k^{FoR2008}\right) \\ 0 & \text{otherwise} \end{cases}$$

$$mapped_i^{FoR2008} = \begin{cases} 1 & \text{if } (\Psi_j^{FoR2008} > 0) \text{ and} \\ & \left(\sum_{k=1}^{|CS^{FoR2008}|} \Psi_k^{FoR2008} = \sum_{k=1}^{|CS^{FoR2020}|} \text{truth}_k^{FoR2020}\right) \\ 0 & \text{otherwise} \end{cases}$$

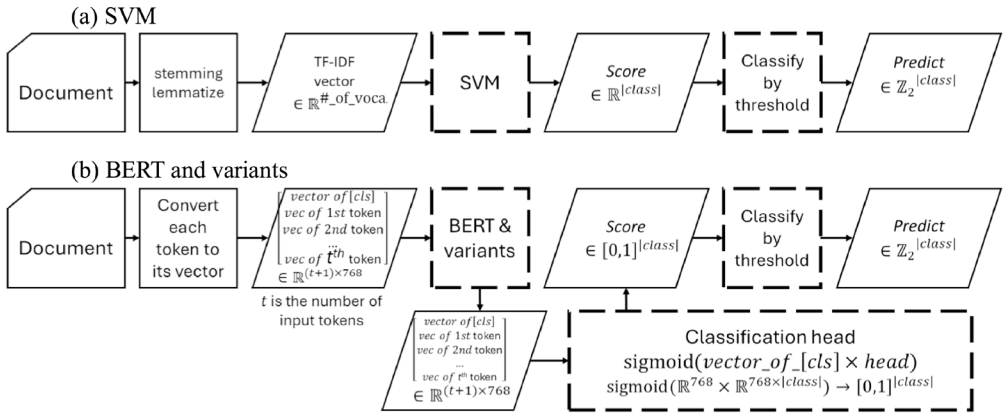
The dataset p08 is derived solely from the $truth^{FoR2008}$ of all documents, whereas m08 incorporates both $truth^{FoR2008}$ and $mapped^{FoR2008}$. In contrast, p20 is built from $truth^{FoR2010}$, while m20 uses both $truth^{FoR2010}$ and $mapped^{FoR2020}$. Each dataset is partitioned into subsets based on three hierarchical levels. By default, the mapping matrix *MAP* is a linear operator that transforms field classes from FoR2008 to FoR2020. The mapping matrix can be across the scheme hierarchy. For example, $MAP \in \mathbb{Z}_2^{|FoR2008 \text{ field classes}| \times |FoR2020 \text{ division classes}|}$ can map a document's classes from the FoR2008 field level to the FoR2020 division level with the above mapping procedure. A class may be possibly-mapped at the field level but definitely-mapped at the division level. If all of a document's classes are definitely mapped at the division level, the document is included in the division-level subset of the mapped dataset. The same criterion applies to the group-level and field-level subsets.

For each dataset, the documents in the field-level subset must not be labeled with any field-level miscellaneous, group-level, or division-level classes. Similarly, the documents in the group-level subset must not be labeled with any group-level miscellaneous, or division-level classes. The field-level miscellaneous classes are viewed as the group-level classes since we observed that some documents labeled with miscellaneous field classes actually refer to the group classes. Each dataset is split into a training set containing 80% of documents and a validation set containing the remaining 20%. All documents in sparse classes, which contain less than five documents, are designated to the training set. Each class containing more than five documents was iteratively sampled to ensure the percentage of the documents in the training set ranged from 76% to 84%. To avoid over-sampling the multi-label documents, the probability of a document to be sampled in the training set was $1 - (1 - 0.8)^{1/k}$, where k is the document's number of classes. The sampling process proceeds sequentially from the field level subset to the division level subset to ensure complete separation between training and validation sets across all three hierarchical levels. Consequently, 179,431 documents are organized into two plain datasets, i.e., p08 and p20, and two mapped datasets, i.e., m08 and m20.

5.2 Classifier

Classifiers are built using SVM, SciBERT, and ModernBERT models. Figure 1 illustrates

Figure 1. Training Approaches



Note. The dashed rectangles are finetuned or trained in this study. $|class|$ denotes the number of classes.

the training approaches. Linear kernel SVM are trained using the Scikit-learn library (https://scikit-learn.org/). Uncased SciBERT (scibert_scivocab_uncased), ModernBERT-base, and ModernBERT-large are downloaded from the Hugging Face (https://huggingface.com/) and finetuned using PyTorch 2.5.1 on an NVIDIA RTX A6000. The model parameters are fully finetuned and only the first token, [CLS], of the output sequence was utilized for our downstream classification task, which was implemented by appending a fully connected layer and a sigmoid layer to the special classification token. The dimension of an output token is 768 for both SciBERT and ModernBERT-base, and 1,024 for ModernBERT-large. The fine-tuning process for BERT models was conducted across varying maximum epochs (1, 2, 4, 8, 16, 24, or 32) with a batch size of 32, employing binary cross-entropy as the loss function, AdamW as the optimizer, and a learning rate of 5e-5. The threshold of each class is determined by iterating over all in-class scores in the training set to maximize the F1 measure of the training set. Classes with scores exceeding their thresholds are selected as predictions. If no scores exceed the thresholds, the class with the highest score is predicted.

6. Result

The initial section covers the preparation task of organizing the datasets. The second section is the evaluation of traditional machine learning and deep learning classifiers. The effect of class size and model parameter size is presented in the final section.

6.1 Dataset profiling

179,431 documents are apportioned into the four datasets as shown in Table 4. The percentage of one-class documents ranges from 49.1% (field-level subset of p20) to 83.5% (division-level p08). 75.0% (field-level p20) to 98.1% (division-level p08) of documents are assigned within 2 classes, whereas 97% (field-level p20) to 99.9% (division-level p08) of documents are assigned within 3 classes. The plain datasets are augmented with class mapping, resulting in an expansion

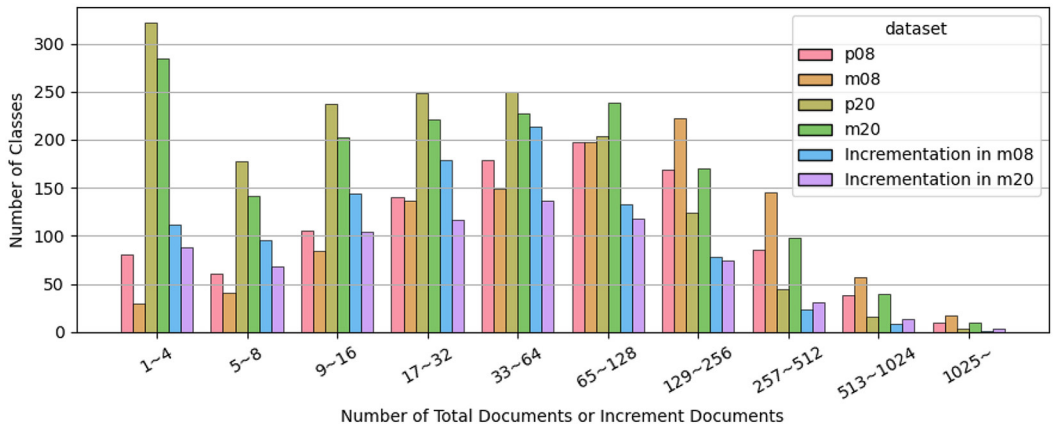
of both the number of documents and the field-level classes. When mapping from FoR2020 to FoR2008, 63% (32,589/51,618) of the p20 documents are mapped into 975 FoR2008 classes. On the contrary, 49% (44,159/89,745) of the p08 documents are mapped into 747 FoR2020 classes. The field-level classes are most affected by data insufficiency, with 37 (m08) to 414 (p20) field classes having too few documents to be sampled in the validation sets. Figure 2 illustrates that the 1~4 document range contains the highest number of field classes for both p20 and m20 datasets. By comparing plain and mapped datasets, the number of classes containing fewer than 64 documents decreases, while the number of classes with more than 65 documents becomes enriched. All 17 newly added classes in the mapped datasets contain fewer than 16 documents, with 9 classes having less than 5 documents. This implies that most of the increase in the number of classes within the validation sets originates from a combination of plain and mapped sources. The mean number of documents per field class increases from 123.7 in p08 to 175.7 in m08, and from 57.4 in p20 to 93.0 in m20. The average document incrementation per mapped class is 58.3 when mapping from FoR2020 to FoR2008, while it reaches 77.7 when mapping from FoR2008 to FoR2020. The testing set strategy is not employed since 35.3% (p20) and 30.1% (m20) of the FoR2020 field classes have less than 10 total documents. The average document length is 210 words, with a median of 191 words. 4,358 (2%) documents exceed 512 words. The deep learning

Table 4. Profile of Datasets

Scheme	Dataset	Level	Number of documents in		# of Classes	# of Class per Doc.
			Training set	Validation set		
ANZSRC 2008 FoR (FoR2008)	p08	Division	98,553	26,600	22	1.19
		Group	93,355	23,416	135	1.32
		Field	74,778	14,967	1,066 (955)	1.47
	m08	Division	138,637	36,720	22	1.25
		Group	127,601	32,139	135	1.39
		Field	102,197	20,137	1,078 (1,041)	1.55
ANZSRC 2020 FoR (FoR2020)	p20	Division	55,877	14,512	23	1.37
		Group	52,211	12,922	189 (187)	1.62
		Field	43,812	7,806	1,627 (1,213)	1.81
	m20	Division	118,967	30,150	23	1.27
		Group	104,321	25,474	189 (187)	1.45
		Field	80,314	15,463	1,632 (1,286)	1.59

Note. The number of classes in the validation set is noted in parentheses if it differs from the number of classes in the training set.

Figure 2. Datasets Distribution



models in our setup support up to 512 tokens and the documents exceeding these token length limits are truncated. By enriching datasets with class mapping, the next section evaluates the classifiers trained on both the plain or mapped datasets.

6.2 Classifier performance

The classification performance of SVM, SciBERT, ModernBERT-base, and ModernBERT-large is listed in Table 5. ModernBERT-large demonstrates superior performance in terms of validation F1 score across nearly all evaluated datasets. SVM with TF-IDF representation is selected for its superior performance compared to experimented traditional machine learning models, which are KNN, Logistic Regression, XGBoost, Linear Classification, Random Forest, Decision Tree, and Naïve Bayes, listed in descending order performance based on the validation macro F1-score. SVM outperforms ModernBERT-base at the field-level subsets. Nevertheless, SVM remains a practical solution for research classification for the online database as of 2023 (Porter et al., 2023). In our experiments, SciBERT performed comparably to BERT-large at the division and group level. Furthermore, field-level classifiers using the original BERT did not demonstrate notable performance. Garcia-Silva and Gomez-Perez (2021) employed 5-fold cross-validation that SciBERT achieved a macro F1 score of 0.838 at the division level, while BERT, SciBERT, and SVMs achieved scores ranging from 0.808 to 0.911 at the group level classes, whereas fastText and GPT-2 were left behind. The relatively inferior macro F1 score in our setting may partly due to the fact that the bibliographic records are from eight repositories, and class labeling may not be inconsistent by various parties. However, SVM falls behind BERT in our study, highlighting a methodological concern for studies that train on the machine-classified records from the online database. Our experiment exhibits that BERT variants outperform all other traditional machine learning methods at the division and group levels. Arhiliuc et al. (2025) reported a macro F1 score of 0.70 for BERT and 0.65 for SVM on 42 FORD

Table 5. Average Precision, Recall, and Macro F1-Score

Level	Dataset	SVM		SciBERT	
		Precision, recall, F1-score of		Precision, recall, F1-score of	
		Training set	Validation set	Training set	Validation set
Division	p08	.893 .923 .908	.661 .677 .668	.803 .775 .788	.727 .698 .711
	m08	.864 .901 .882	.657 .681 .668	.796 .773 .784	.725 .708 .715
	p20	.915 .933 .924	.673 .667 .669	.999 .999 .999	.706 .710 .707
	m20	.882 .905 .893	.686 .686 .685	.802 .787 .793	.737 .720 .727
Group	p08	.935 .962 .948	.527 .463 .481	.995 .987 .991	.512 .522 .511
	m08	.912 .942 .926	.535 .485 .502	.998 .996 .997	.538 .533 .530
	p20	.962 .976 .969	.540 .433 .471	.973 .967 .968	.514 .491 .494
	m20	.935 .958 .946	.530 .449 .475	.993 .983 .987	.533 .508 .512
Field	p08	.987 .995 .991	.383 .278 .303	.977 .966 .970	.368 .310 **.318§**
	m08	.979 .990 .985	.394 .286 .312	.945 .931 .935	.365 .320 .325
	p20	.995 .998 .996	.341 .244 .265	.931 .918 .918	.301 .247 .252
	m20	.990 .996 .993	.346 .248 .271	.944 .935 .933	.329 .277 .284
		ModernBERT-base		ModernBERT-large	
Division	p08	.786 .773 .778	.716 .698 .706	.793 .780 .786	.723 .712 **.716§**
	m08	.781 .769 .774	.722 .708 .713	.799 .775 .785	.735 .714 **.723§**
	p20	.771 .763 .766	.709 .695 .701	.999 .999 .999	.720 .714 **.716§**
	m20	.780 .769 .773	.729 .710 .718	.794 .787 .789	.745 .727 **.734§**
Group	p08	.676 .675 .671	.501 .499 .494	.593 .624 .602	.516 .551 **.524§**
	m08	.700 .691 .693	.545 .512 .520	.625 .637 .628	.545 .556 **.547§**
	p20	.852 .956 .900	.511 .462 .474	.975 .978 .976	.557 .484 **.507§**
	m20	.983 .981 .982	.531 .480 .496	.563 .591 .569	.524 .527 **.518§**
Field	p08	.876 .855 .857	.331 .292 .292	.970 .978 .972	.366 .303 .314
	m08	.784 .748 .754	.323 .301 .293	.799 .804 .795	.386 .343 **.345§**
	p20	.932 .915 .916	.306 .251 .256	.938 .933 .930	.327 .277 **.279§**
	m20	.872 .834 .840	.307 .266 .267	.788 .788 .769	.353 .313 **.311§**

Note. The bold entries suffixed by **§** are the largest macro F1 score of the validation sets.

classes, with metrics similar to the division level metrics in our study. Wu et al. (2023) trained with 76 to 600 records per division class and reported accuracy values between 0.60 to 0.70. In contrast, our study achieves all accuracies exceeding 0.941. In summary, ModernBERT-large demonstrates superior performance compared to alternative architectures across the majority of datasets and hierarchical classification levels, with the sole exception being the p08 dataset where SciBERT maintains a marginal advantage. The consequence of class mapping using ModernBERT-large is analyzed in greater detail in the following sections.

ModernBERT requires fewer or an equivalent number of epochs to achieve optimal validation F1 scores on mapped datasets compared to their plain counterparts, potentially reducing overall training time requirements. Table 6 presents the training epochs, maximum epoch, and minutes per epoch. Training epoch duration is primarily influenced by dataset volume, model architecture, and model parameter size. As shown in Table 6, the training time per epoch is proportional to the dataset volume. When accounting for model architecture, ModernBERT-base exhibits approximately half the per-epoch training time of SciBERT despite having a comparable parameter count. Similarly, ModernBERT-large achieves nearly a twofold reduction in training time per epoch compared to BERT-large. The number of training epochs is proportional to the number of classes. Lower levels, which correspond to a greater number of classes, generally require more training epochs. As to the total training time, the division and group subsets of m20, as well as the field subset of m08, exhibit fewer total time comparing to their plain counterparts. While class mapping increases both

Table 6. BERT Training Time and Epoch

		SciBERT (110 Million)	ModernBert-base (149M)	ModernBert-large (395M)	BERT-large (336M)
Division	p08	2nd epoch/2 Max. epoch (20 mins per epoch)	2/2 (11)	2/2 (27)	(61 mins per epoch)
	m08	2/2 (28)	2/2 (15)	2/2 (39)	(86)
	p20	8/8 (11)	2/2 (7)	7/8 (16)	(34)
	m20	2/2 (24)	2/2 (13)	2/2 (33)	(73)
Group	p08	12/16 (19)	3/4 (10)	2/2 (25)	(57)
	m08	13/16 (26)	3/4 (14)	2/2 (35)	(78)
	p20	11/24 (10)	10/16 (6)	6/16 (15)	(32)
	m20	12/16 (21)	8/24 (12)	2/2 (30)	(64)
Field	p08	20/32 (15)	7/24 (8)	8/16 (20)	(45)
	m08	16/32 (21)	6/24 (11)	5/8 (28)	(62)
	p20	25/32 (9)	9/32 (5)	8/16 (13)	(26)
	m20	21/32 (16)	7/16 (9)	6/16 (23)	(49)

dataset size and per-epoch training time, it can, in some cases, reduce the overall time required to achieve optimal performance.

Class mapping enhances the validation F1 scores as illustrated in Figure 3. At the division level, the F1 score improvement between p08 and m08 is at most 0.007, whereas the improvement between p20 and m20 is at least 0.016. This may be due to the dataset size increasing by 112% (78,728 documents) in m20, compared to only a 40% (50,204 documents) increase in m08. FoR2020, as a newer scheme, has relatively fewer training documents, which can be supplemented through class mapping. In contrast, p08 already has a sufficient number of documents, making class mapping less effective in improving classifier performance. While focusing on the best-performing model, the validation F1-score improvements of ModernBERT-large in FoR2008 increased by 1.0%, 4.4%, and 9.9% at division, group, and field levels respectively. On the contrary, the enhancements in FoR2020 are 2.5%, 2.2%, and 11.5% respectively.

6.3 Class size and model parameter size

Class size, which is the number of documents in a class, demonstrates a moderate positive correlation with the validation F1 score at the field level. Specifically, Pearson correlation coefficients are 0.36 for p08, 0.33 for m08, 0.35 for p20, and 0.31 for m20 in the ModernBERT-large model. This implies that larger document quantities generally lead to higher performance. Figure 4 illustrates the relationship between class size and validation F1 scores. Classes containing fewer than 16 documents consistently show extremely poor performance, with F1 scores predominantly at 0, resulting in both lower and upper quartiles also at 0. The average F1 score consistently increases with class size, showing progressive improvement as class size grows, with the exception of the bucket exceeding 1,025 documents. The number of classes exceeding 1,025 documents is rare that 10 classes in p08, 17 in m08, 3 in p20, and 10 in m20. High-volume classes with F1-scores below 0.5 are listed in Table 7. Most augmented classes are improved, except for 050202 and 410406, which

Figure 3. Growth of Validation F1 Score

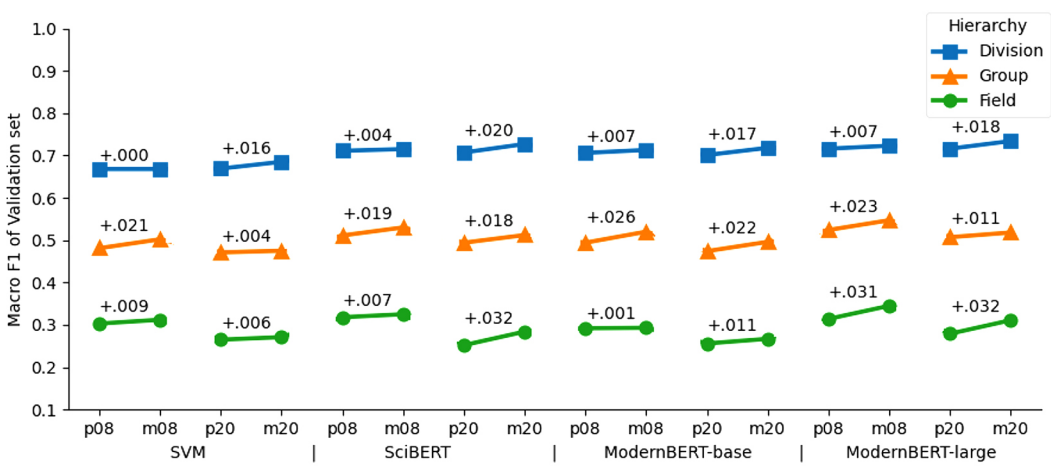
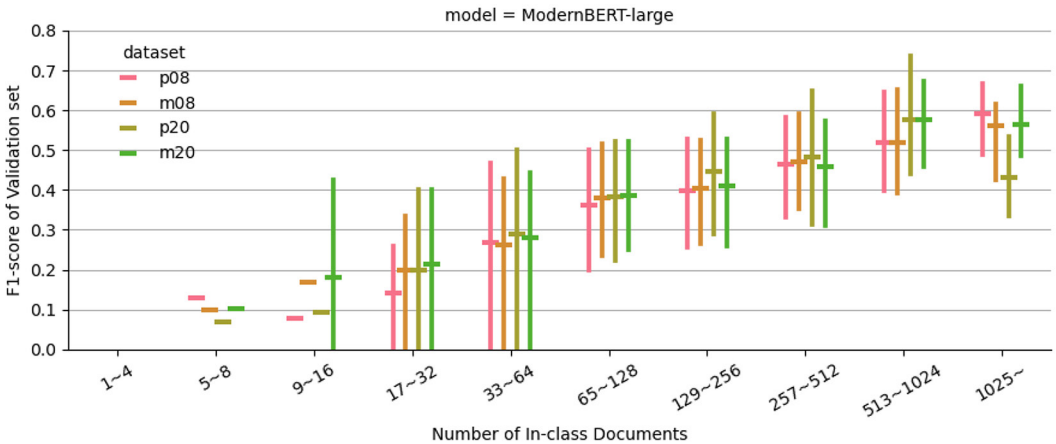


Figure 4. Document Size and Validation F1 Score



Note. The horizontal lines denote the average F1; the vertical lines denote the range between the lower quartile and the upper quartile.

Table 7. Validation F1 Scores and Document Counts of Selected High-volume Classes

	plain dataset	mapped dataset
050209 Natural resource management	0.244 (970)	0.313 (1,269)
410406 Natural resource management	0.311 (483)	0.236 (1,108)
050205 Environmental management	0.318 (1,412)	0.296 (1,412)
410404 Environmental management	0.219 (1,047)	0.252 (1,877)
050202 Conservation and biodiversity	0.475 (2,019)	0.401 (2,697)
410401 Conservation and biodiversity	0.393 (1,012)	0.470 (2,288)
130103 Higher education	0.276 (743)	0.425 (1,967)
390303 Higher education	0.447 (1,578)	0.534 (1,914)
160104 Social and cultural anthropology	0.412 (893)	0.414 (1,075)
440107 Social and cultural anthropology	0.363 (220)	0.304 (220)
111706 Epidemiology	0.476 (1,647)	0.524 (2,126)

are augmented with worse-performed documents. These findings suggest that low-quality data can negatively impact model performance. Overall, the validation F1-score tends to increase proportionally with the class size.

Class mapping improves the average F1-score not only for the augmented classes but also for the non-augmented ones in ModernBERT-large as Table 8 demonstrates. In ModernBERT-base, only the F1-scores of the augmented classes

Table 8. Average F1 Score of Augmented and Non-augmented Classes

		P08	m08	p20	m20
ModernBERT-base	Augmented class	0.292	0.297	0.285	0.308
	Non-augmented class	0.217	0.212	0.207	0.204
ModernBERT-large	Augmented class	0.306	0.348	0.303	0.353
	Non-augmented class	0.197	0.260	0.207	0.239

showed slight improvements of 0.005 and 0.023, while the performance of the non-augmented classes declined by 0.003 and 0.005. In contrast, ModernBERT-large demonstrated notable gains, with non-augmented classes improving by 0.063 and 0.032, and augmented classes achieving increases of 0.042 and 0.050. This suggests that the model size plays a role in enhancing the effectiveness of mapping.

The largest groups in Figure 5 are the zero F1 classes, which continues to impede the overall performance of field-level classifiers on validation sets. The larger model exhibits more zero F1 classes in plain datasets compared to the base model. However, mapped datasets show a significant reduction in these zero F1 classes, particularly when using the large model. Tables 9 and 10 detail the transition of classes between F1-score groups by ModernBERT-large. Zero F1 classes numbered 220 in p08, 287 in m08, 400 in p20, and 510 in m20. A transition matrix is defined as $T \in \mathbb{N}^{12\times12}$. $T_{i,j}$ denotes the number of classes which are categorized into the group i of the plain dataset and into the group j of the mapped dataset. When transitions occur, the F1 score between 0.4 and 0.5 ($T_{7,7}$) serves as a critical inflection range. Classes with F1 scores below 0.4 predominantly shift to higher performance groups, as $\Sigma_{j=1}^{k-1} T_{k,j} < \Sigma_{j=k+1}^{12} T_{k,j}$, $\forall k < 7$, while those with F1 scores exceeding 0.5 predominantly shift to lower performance groups, as $\Sigma_{j=1}^{k-1} T_{k,j} > \Sigma_{j=k+1}^{12} T_{k,j}$, $\forall k > 7$. The steady state derived from the Markov chain analysis, which is $s \in \mathbb{R}^{12}$ such that $s \times U = s$, *where* $U_{i,j} = T_{i,j}/\Sigma_{j=1}^{12} T_{i,j}$ and displayed in the bottom row of the tables, reveals that the performance group with F1 scores between 0.4 and 0.5 will emerge as the dominant category through iterative class mapping process. Although 25% of the classes in m20 have zero F1 scores, this ratio is expected to decrease to 16% through additional class mapping. By summing the products of each steady-state share and the midpoint of its corresponding range, the estimated validation F1 score for ModernBERT-large is 0.452 for FoR2008 and 0.363 for FoR2020. Since FoR2020 is a presently mandated revision that will incorporate new records, the macro F1 score is expected to be improved with the increasing document counts in the future.

7. Conclusion

This study demonstrates the effectiveness of class mapping as a data fusion technique to improve machine learning-based research classification. SVM, SciBERT, ModernBERT-base, and ModernBERT-large are used to train

Figure 5. Classes Group by the Validation F1-score

classifiers for the ANZSRC 2008 FoR (FoR2008) and ANZSRC 2020 FoR (FoR2020) research classification schemes across three hierarchical levels. By leveraging the definitely-mapped relations between FoR2008 and FoR2020, 63% of FoR2020 documents are incorporated into the mapped dataset of FoR2008, while 49% of FoR2008 documents are integrated into the mapped dataset of FoR2020. Although the mapped datasets still contain many low-volume classes, particularly in FoR2020, class mapping substantially increases both the number of documents and the range of represented classes, supporting improved classifier performance. Among the evaluated models, ModernBERT-large consistently delivers strong performance, particularly on mapped datasets. SVM remains a practical baseline, offering competitive results comparable to base-sized BERT variants at the field level. Porter et al. (2023) introduced recategorization strategies, including class mapping, but did not provide empirical validation. In contrast, our study provides concrete metrics to evaluate the effectiveness of class mapping, addressing a gap in the literature where research classification involving field classes has not been previously explored (Wu et al., 2021). Consistent performance gained across all three scheme levels

Table 9. Transition of Validation F1-score from p08 to m08

p08 \ m08	Excluded	[0,0]	(.0,.1]	(.1,.2]	(.2,.3]	(.3,.4]	(.4,.5]	(.5,.6]	(.6,.7]	(.7,.8]	(.8,.9]	(.9,1.0]	Total
Exclud.	35	58			4	2	9	2	5	3	2	3	123
[0,0]	2	140	5	25	35	18	27	9	14	3	5	4	287
(.0,.1]			1	3	5	1	1						11
(.1,.2]			4	19	18	18	10	1					70
(.2,.3]		4	1	20	30	26	20	13	4	1			119
(.3,.4]		4	1	9	21	22	39	4	4				104
(.4,.5]		6		3	18	22	50	31	13	2			145
(.5,.6]		1		2	3	8	20	40	16	3			93
(.6,.7]		3		1	1	3	2	13	31	12	3		69
(.7,.8]					1			4	8	15	3		31
(.8,.9]							1	2	3	4	4		14
(.9,1.0]		4				1	2			3	1	1	12
Total	37	220	12	82	136	121	181	119	98	46	18	8	1,078
Share	.03	.20	.01	.08	.13	.11	.17	.11	.09	.04	.02	.01	
Markov	.00	.05	.01	.06	.11	.12	.19	.18	.16	.09	.02	.00	

Table 10. Transition of Validation F1-score from p20 to m20

p20 \ m20	Excluded	[0,0]	(.0,.1]	(.1,.2]	(.2,.3]	(.3,.4]	(.4,.5]	(.5,.6]	(.6,.7]	(.7,.8]	(.8,.9]	(.9,1.0]	Total
Exclud.	334	59		1	1	3	8	1	6	2		4	419
[0,0]	12	279	5	22	60	30	46	10	28	5	5	8	510
(.0,.1]				2	2	1							5
(.1,.2]		7	5	11	15	16	9	3	3	1			70
(.2,.3]		13	4	13	31	18	32	6	2				119
(.3,.4]		9	2	8	20	21	26	8	4	1	1		100
(.4,.5]		19	1	9	18	24	52	21	12	2	3	1	162
(.5,.6]		2		4	8	7	16	25	16	3	1		82
(.6,.7]		7		1	3	10	19	16	17	6	6	2	87
(.7,.8]					1	1	3	6	11	6	5	1	34
(.8,.9]		2			1		1	2	6	11	9		32
(.9,1.0]		3					2		1	1	2	3	12
Total	346	400	17	71	160	131	214	98	106	38	32	19	1,632
Share	.21	.25	.01	.04	.10	.08	.13	.06	.06	.02	.02	.01	
Markov	.02	.16	.01	.07	.14	.13	.21	.11	.09	.03	.03	.01	

in four models is exhibited by comparisons of validation F1 scores between plain and mapped datasets. Class size shows a moderate correlation with the validation F1 score, and some of classes are augmented via class mapping. On the contrary, the performance of non-augmented classes is also generally improved in ModernBERT-large but not enhanced in ModernBERT-base. The advantage of class mapping on non-augmented classes is the emergent ability, which emerges only in larger models (Wei et al. 2022). This study recommends leveraging a larger LLM in conjunction with class mapping to enhance the effectiveness of research classification models. While larger models and more training data increase the time per epoch, they require fewer epochs to reach optimal performance and sometimes reduced total training time. Although decoder-based Transformer architectures remain dominant as of 2025, this study advocates for the development of larger encoder models to enable more effective research classification in the future. Inconsistent classification is the most significant limitation encountered, as it adversely impacts the performance. For example, Class 130103 Higher Education erroneously includes articles based solely on studies conducted in college or university settings, which do not accurately reflect the intended scope of the class. This study employs only the definitely-mapped class relations for dataset augmentation, while the possibly-mapped relations may be explored in future work, such as ensemble modeling or contrastive learning.

References

Arhiliuc, C., Guns, R., Daelemans, W., & Engels, T. C. (2025). Journal article classification using abstracts: A comparison of classical and transformer-based machine learning methods. *Scientometrics, 130*, 313-342. https://doi.org/10.1007/s11192-024-05217-7

Australian Bureau of Statistics. (2020a). *ANZSRC 2020 correspondence to ANZSRC 2008* [Data Set]. Retrieved January 28, 2024, from https://pse.is/7tsx7e

Australian Bureau of Statistics. (2020b). *Australian and New Zealand Standard Research Classification (ANZSRC)*. Retrieved January 28, 2024, from https://pse.is/7tsx95

Australian Research Council. (n.d.). *Classification codes: FoR, RFCD, SEO and ANZSIC codes*. Retrieved March 22, 2021, from https://www.arc.gov.au/grants/grant-application/classification-codes-rfcd-seo-and-anzsic-codes

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3615-3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the*

Association for Computational Linguistics, 5, 135-146. https://doi.org/10.1162/tacl_a_00051

Bornmann, L. (2018). Field classification of publications in Dimensions: A first case study testing its reliability and validity. *Scientometrics, 117*, 637-640. https://doi.org/10.1007/s11192-018-2855-y

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*, 2493-2537. https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf

Commonwealth of Australia and New Zealand. (2020). *Outcomes paper: Australian and New Zealand Standard Research Classification review 2019*. https://www.arc.gov.au/sites/default/files/anzsrc_review_outcomes_paper_v1.1.pdf

Dahlberg, I. (1993). Knowledge organization: Its scope and possibilities. *Knowledge Organization, 20*(4), 211-222. https://doi.org/10.5771/0943-7444-1993-4-211

Dahlberg, I. (1998). Classification structure principles: Investigations, experiences and conclusions. In W. Mustafa el Hadi, J. Maniez, & S. A. Pollitt (Eds.), *Structures and relations in knowledge organization: Proceedings of the 5th international ISKO conference* (Advanced in Knowledge Organization, Vol. 6, pp. 80-88). Ergon Verlag.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Digital Science and Research Solutions. (2022a, October 6). *What is the background behind the Fields of Research (FoR) classification system?* Dimensions. https://pse.is/7tszew

Digital Science and Research Solutions. (2022b, October 6). *Which research categories and classification schemes are available in Dimensions?* Dimensions. https://pse.is/7tszfz

Garcia-Silva, A., & Gomez-Perez, J. M. (2021). Classifying scientific publications with BERT: Is self-attention a feature selection method? In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Advances in information retrieval* (pp. 161-175). https://doi.org/10.1007/978-3-030-72113-8_11

Hammarfelt, B. (2020). Discipline. *Knowledge Organization, 47*(3), 244-256. https://doi.org/10.5771/0943-7444-2020-3-244

Hider, P., & Coe, M. (2022). Academic disciplines in the context of library classification: Mapping university faculty structures to the DDC and LCC schemes. *Cataloging & Classification Quarterly, 60*(2), 194-213. https://doi.org/10.1080/01639374.2022.2040675

Hjørland, B., & Gnoli, C. (Eds.). (2022, June 15). Research classification system. In *ISKO encyclopedia of knowledge organization*. https://www.isko.org/cyclo/research

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning, 32*(2), 1188-1196. https://proceedings.mlr.press/v32/le14.html

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Legendere, A. (2019). The development of the Canadian Research and Development Classification. *Knowledge Organization, 46*(5), 371-379. https://doi.org/10.5771/0943-7444-2019-5-371

Macauley, P., Evans, T., & Pearson, M. (2011). *Classifying Australian PhD bibliographic thesis records by ANZSRC field of research codes*. Australian Research Council Research Excellence Branch. http://hdl.handle.net/10536/DRO/DU:30036705

Meo-Evoli, L., Negrini, G., & Farnesi, T. (1998). ICC and ICS: Comparison and relations between two systems based on different principles. In W. Mustafa el Hadi, J. Maniez, & S. A. Pollitt (Eds.), *Structures and relations in knowledge organization: Proceedings of the 5th international ISKO conference* (Advanced in Knowledge Organization, Vol. 6, pp. 228-236). Ergon Verlag.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, May 2-4). *Efficient estimation of word representations in vector space* [Conference Workshop Poster Session Abstract]. International Conference on Learning Representations 2013, Scottsdale, AZ, United States. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111-3119. https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Organisation for Economic Co-operation and Development. (2015). *Frascati manual 2015: Guidelines for collecting and reporting data on research and experimental development* (The Measurement of Scientific, Technological and Innovation Activities). OECD Publishing. https://doi.org/10.1787/9789264239012-en

Pasin, M. (Ed.). (2017). *Data Release: 2017Q1 · springernature/scigraph*. Github. Retrieved March 31, 2024, from https://github.com/springernature/scigraph/wiki/Data-Release:-2017Q1

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human*

language technologies (Vol. 1, pp. 2227-2237). https://doi.org/10.18653/v1/N18-1202

Porter, S. J., Hawizy, L., & Hook, D. W. (2023). Recategorising research: Mapping from FoR 2008 to FoR 2020 in Dimensions. *Quantitative Science Studies, 4*(1), 127-143. https://doi.org/10.1162/qss_a_00244

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47. https://doi.org/10.1145/505282.505283

United Nations Educational, Scientific and Cultural Organization. (2015). *International Standard Classification of Education: Fields of education and training 2013 (ISCED-F 2013) – Detailed field descriptions*. UNESCO Institute for Statistics. http://doi.org/10.15220/978-92-9189-179-5-en

Vancauwenbergh, S., & Poelmans, H. (2019). The Flemish Research Discipline Classification Standard: A practical approach. *Knowledge Organization, 46*(5), 354-363. https://doi.org/10.5771/0943-7444-2019-5-354

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998-6008. https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Warner B., Chaffin A., Clavié B., Weller O., Hallström O., Taghadouini S., Gallagher A., Biswas R., Ladhak F., Aarsen T., Cooper N., Adams G., Howard J., & Poli I. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv*. https://doi.org/10.48550/arXiv.2412.13663

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://openreview.net/pdf?id=yzkSU5zdwD

Wu, M., Brandhorst, H., Marinescu, M.-C., Lopez, J. M., Hlava, M., & Busch, J. (2023). Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence, 5*(1), 122-138. https://doi.org/10.1162/dint_a_00162

Wu, M., Liu, Y.-H., Brownlee, R., & Zhang, X. (2021). Evaluating utility and automatic classification of subject metadata from Research Data Australia. *Knowledge Organization, 48*(3), 219-230. https://doi.org/10.5771/0943-7444-2021-3-219

Zeng, M. L. (2019). Interoperability. *Knowledge Organization, 46*(2), 122-146. https://doi.org/10.5771/0943-7444-2019-2-122

Zhang, L., Sun, B., Shu, F., & Huang, Y. (2022). Comparing paper level classifications across different methods and systems: An investigation of Nature publications. *Scientometrics, 127*(12), 7633-7651. https://doi.org/10.1007/s11192-022-04352-3

Zhang, S., Wu, M., & Zhang, X. (2023). Utilising a large language model to annotate subject metadata: A case study in an Australian national research data catalogue. *arXiv*. https://doi.org/10.48550/arXiv.2310.11318

(Received: 2024/12/7; Accepted: 2025/5/8)

圖書資訊學刊　第23卷第2期　（2025.12）　頁119-143
https://doi.org/10.6182/jlis.202512_23(2).119

應用類別對照之資料融合方法於機器學習研究分類

Exploring Class Mapping as Data Fusion Technique in Machine Learning for Research Classification

黃建智[1]　陳光華[2]

Chien-Chih Huang[1], Kuang-Hua Chen[2]

摘　要

訓練機器學習分類模型須充足且高品質的資料，本研究探討類別對照作為資料融合策略，發展研究分類之機器學習模型，以2008年版與2020年版之澳洲與紐西蘭標準研究分類表為研究標的，從8家機構典藏系統蒐集179,431筆已分類文件，對二版本分類表分別建立原始資料集，及以類別對照方式擴增之資料集。結果顯示49%的2008年版文件可明確對應至2020年版，反之則為63%。進一步以SVM、SciBERT、ModernBERT-base與ModernBERT-large建立分類模型，相較僅採用原始資料集，各模型經擴增資料集訓練後，分類效能均獲改進；以ModernBERT-large表現最為顯著，其大類層級提升1.0%或2.5%，中類層級增益4.4%或2.2%，小類層級改善9.9%或11.5%，未擴增之類別亦提高32.0%或15.5%。整體而言，類別對照可用於擴展訓練資料，提升自動研究分類效能。

關鍵字：互通性、概念間對應、機器學習、研究分類

[1,2] 國立臺灣大學圖書資訊學系

Department of Library and Information Science, National Taiwan University, Taipei, Taiwan

* 通訊作者Corresponding Author: 陳光華Kuang-Hua Chen, E-mail: khchen@ntu.edu.tw

註：本中文摘要由作者提供。

以APA格式引用本文：Huang, C.-C., & Chen, K.-H. (2025). Exploring class mapping as data fusion technique in machine learning for research classification. *Journal of Library and Information Studies, 23*(2), 119-143. https://doi.org/10.6182/jlis.202512_23(2).119

以Chicago格式引用本文：Huang, Chien-Chih, and Kuang-Hua Chen. “Exploring Class Mapping as Data Fusion Technique in Machine Learning for Research Classification.” *Journal of Library and Information Studies* 23, no. 2 (2025): 119-143. https://doi.org/10.6182/jlis.202512_23(2).119