Early Modern Book Catalogues and Multilingualism: Identifying Multilingual Texts and Translations using Titles

Yann Ryan[1], Margherita Fantoli[1]

[1]Faculty of Arts, KU Leuven, Blijde-Inkomststraat 21, 3000 Leuven, Belgium

Abstract

With this paper we aim to assess whether Early Modern book titles can be exploited to track two aspects of multilingualism in book publishing: publications featuring multiple languages and the distinction between editions of works in their original language and in translation. To this scope we leverage the manually annotated language information available in two book catalogs: the Collectio Academica Antiqua, recording publications of scholars of the Old University of Leuven (1425-1797) and a subset of the Eighteenth Century Collections Online, namely publications of Ancient Greek and Latin works. We evaluate three different approaches: we train a simple tf-idf based support vector classifier, we fine-tune a multilingual transformer model (BERT) and we use a few-shot approach with a pre-trained sentence transformer model. In order to get a better understanding of the results, we make use of SHAP, a library for explaining the output of any machine Learning model. We conclude that while the few-shot prediction is not currently usable for this task, the tf-idf approach and BERT fine-tuning are comparable and both usable. BERT shows better results for the task of identifying translations and when generalizing across different datasets.

Keywords

multilingualism, metadata, transformer models, few-shot classification, library catalogues,

1. Introduction

Metadata catalogues, particularly library catalogues, are increasingly valuable for reconstructing the cultural and intellectual life of the past [33, 18, 30, 27, 19]. These catalogues provide insights into both cultural artefacts and the actors behind the publishing industry, often spanning vast temporal and spatial ranges. Widely implemented metadata schemes such as MARC21[1] and Dublin Core[2] facilitate large-scale mining of these resources. The manual creation of catalogues, relying on experts familiar with the epoch and place covered, as well as cataloguing best practices, ensures their reliability as data sources.

In this paper, we aim at investigating whether machine learning and Large Language Models can support the labelling of Early Modern book records in relation to language. Specifically, we explore the use of titles to identify multilingual publications and distinguish between works published in their original language and those translated. The full titles recorded in several

CHR 2024: Computational Humanities Research Conference, Aarhus, Denmark, December 4-6, 2024

yann.ryan@kuleuven.be (Y. Ryan); margherita.fantoli@kuleuven.be (M. Fantoli)

0000-0003-1878-4838 (Y. Ryan); 0000-0003-1878-4838 (M. Fantoli)

[1]https://www.loc.gov/marc/bibliographic/

[2]https://www.dublincore.org/

catalogues of Early Modern books are highly informative regarding the linguistic form of the book's content: they may mention the translator, the language in which the text is printed, and the language from which the text is translated. A typical is example is provided by the title 'A poetical translation of the works of Horace: with the original text, and critical notes collected from his best Latin and French commentators. By the Rev. d Mr. Philip Francis. In four volumes.'. This paper aims to answer three research questions:

- RQ1: Do the titles recorded in catalogues of Early Modern books contain sufficient information to predict if they were multilingual or monolingual, and printed in the original language or translated?
- RQ2: Which approach yields the best results: a simple tf-idf classifier, training a Large Language Model, or adopting a few-shot approach?
- RQ3: Given the heterogeneity of Early Modern publications, can models trained on one dataset yield satisfactory results on others? Does the diversification of training data improve the results on the datasets analyzed?

The work is structured as follows: in Section 2, we discuss the importance of multilingualism for Early Modern studies and the current possibilities for automatic language information extraction. Section 3 introduces the two datasets used in this experiment.[3] In Section 4, we describe the tasks (Section 4.1) and models (Section 4.2) employed. Finally, Sections 5 and 6 present the results and discuss the potential of this approach.

2. Related work

Early Modern Europe was marked by multilingualism. As Latin's dominance as the *lingua franca* waned, vernacular languages began to emerge in scientific and literary production. This shift influenced various practices in the printed press, drawing interest from linguistics, book history, literary studies, and translation studies [2]. A key focus is the reception of classical texts. During Humanism and the Renaissance, Ancient Greek and Latin gained prominence, and, on the one hand, reading original works became central to humanistic education [22, 17]. On the other hand, this interest led to significant translation efforts, impacting the cultural landscape [4, 11, 20].

This study examines two datasets reflecting aspects of Early Modern multilingualism: the diverse linguistic environment of the Low Countries and the evolving practice of printing classical authors in England. The Low Countries, a multilingual hub due to their political situation [13, 38], saw significant scholarly activity around the Old University of Leuven, captured by the catalog *Collection Academica Antiqua* (CAA). The CAA features several Ancient Greek and Latin authors, reflecting the high value placed on classics in the Low Countries' learned society, as exemplified by the curriculum of the *Collegium Trilingue* [15, 14, 6]. In England, we focus on the printing of Classics in the eighteenth century. The influence of Ancient Greek and Latin on Grammar School curricula and the role of translations in circulating classics have been well-documented [39, 3, 41]. This resulted in multilingual publications recorded in catalogs such

[3]The data and the code are available at: https://github.com/mfantoli/CHR2024_multilingualism.

as the English Short Title Catalog (ESTC) and Eighteenth Century Collections Online (ECCO), the latter used in this study. More details are provided in Section 3.

Our work utilizes long titles of Early Modern books to annotate their linguistic characteristics. Book titles have been leveraged for metadata enrichment and large-scale analysis in several studies: from the decline of the average length of modern British novel titles [25], to genre classification [26],[4] and topic modeling (two examples based on art catalogs are [10, 5]). Recent experiments have leveraged language and multimodal models to semantically enrich metadata sets [40, 1, 24, 31]. In this paper, we assess whether titles can be used to track multilingualism phenomena in a catalogue (i.e., to enrich metadata with specific language information). As noted by Hatzel, Stiemer, Biemann, and Gius [12], traditional, feature-based machine learning approaches are still widely applied in the Humanities. Hence, we compare a tf-idf-based classifier with the performance of Large Language Models (LLMs) [37] (here, BERT [8]), particularly trained for multilingual sentence classification. Transformer-based LLMs are increasingly used for annotation and to enrich metadata or analyse historical text collections, for example to predict the year of publication from text [42], or to investigate genre within books [28]. The availability of multilingual and historical text models, through easy-to-use APIs such as HuggingFace, means that the potential for such models to enhance research or augment our bibliographic understanding of large collections has greatly increased in recent years. Given the high resource cost of fine-tuning LLMs, we also test a few-shot approach for the same task, where only a few examples are used to tune the model (see Section 4.2).

We aim to achieve two objectives: label a work as multilingual or monolingual and identify whether it is printed in the original language or translated. These tasks, while related to language identification [16], are tailored to Early Modern book history: a title may be monolingual but indicate a multilingual work, and identifying the title's language alone is insufficient to determine if it is a translation or an original edition. The presence of multiple languages in metadata sets has already been recognized as a major challenge in metadata processing [23].

3. Data

The present study relies on two datasets: the CAA[5] from KU Leuven, and a version of Eighteenth Century Collections Online (ECCO)[6] manually enriched by a group of students. The CAA is curated by the Special Collections of KU Leuven Libraries and comprises books related to the Old University of Leuven (1425-1797), mostly of scholars that, at a certain point of their career, were affiliated to this university. The CAA version used for this study (exported on 28 July 2023) comprises 3660 holdings, each of them described in MARC XML records. ECCO is a digital database assembled by Gale and stores the (OCRed) full text of a collection of 184,536 titles published in the eighteenth century. Within this collection, we identified the set of classical publications, as those authored by Ancient Greek or Latin authors living before the sixth cen-

[4]Enriching metadata based on book titles is also of interest to GLAM institutions, as demonstrated by a recent experiment on British Library data, https://living-with-machines.github.io/genre-classification/01_BL_fiction_non_fiction.html

[5]https://dial.uclouvain.be/digitization/en/digital-collection/old-academic-collection.

[6]https://www.gale.com/primary-sources/eighteenth-century-collections-online.

language pair	# CAA	language pair	#ECCO
lat|grc	95	lat|eng	876
fre|lat	32	grc|lat	648
lat|heb	16	lat|fre	27
ita | lat	13	grc|lat|eng	31
dut|fre	12	lat|fre|eng	8

Table 1: Most attested language combinations in multilingual works of CAA and ECCO-classics

tury.[7] The total number of classical editions amounts to 5237 rows. We refer to this dataset as ECCO-classics. These two datasets are chosen because of their meticulous language annotation, their partial chronological overlap, the shared presence of classics (several classical works were printed in Early Modern Flanders, and feature in the CAA),[8] but also clear differences in terms of languages included and cultural and geographical background: these characteristics make them useful sets for comparing the capacities of generalization of the different approaches.

3.1. Linguistic annotation

Both datasets have been manually annotated with respect to language. The MARC21 metadata schema includes a specific code for language annotation (041), further specified by several subfields, two of which are used in the CAA: 'a' indicating the language of the record, and 'h', indicating the original language. Hence, **multilingual works** are those including several 'a' codes, regardless of the presence of a 'h' code. **Monolingual works** include only one 'a' code. Within the monolingual works, some also include an 'h' code, which is noted when the original is different from the language of the edition. We speak of **monolingual edition** if no 'h' code is recorded, and **monolingual translation** if it is recorded (and is consequently different from the 'a' code). In fact, monolingual translations are usually works translated into a single target language and published without the original text. We include only monolingual works for identifying translations, because for multilingual works it is hard to single out the function of the different target languages and be sure that one of them is used for translation.

An example of multilingual work in the CAA is represented by 'Les dialogvesde Iean Loys Vives, traduits de Latin en François pour l'exercice des deux langues .../Les dialogues de Jean Loys Vives', which is labeled as French and Latin. Table 1 lists the most frequently attested language combinations for multilingual works in the CAA.

The 'Histoire de Notre-Dame de Hale,par Juste Lipse ... Traduit du latin, & augmentée de plusieurs merveilles, venues en lumière depuis la mort de l'auteur' is the title of a work labeled as monolingual translation. Table 2 shows the most frequent pairs of original and target languages in the CAA. As both Table 1 and 2 demonstrate, translation of the classical languages (Ancient Greek and Latin) plays a central role in the multilingualism of the academic production.

[7] More information on the identification of classical authors is provided in [9].

[8] We haven't counted the exact number of classical works in the CAA, but, as an example, there are at least five editions of Homer, more than 10 editions of Cicero, etc.

source-target languages	# CAA	source-target languages	# ECCO
lat-dut	51	grc-eng	1198
lat-fre	34	lat-eng	926
fre-dutch	11	grc-lat	11
lat-ger	11	grc-fre	26

Table 2: Most attested language combinations in monolingual translations of CAA and ECCO-classics

dataset	monolingual	multilingual	monolingual ed.	monolingual transl.
CAA	3466	194	3291	175
balanced CAA monolingual	200	194	not used	not used
balanced CAA translation	not used	not used	350	175
ECCO-classics	550	1765	1156	609
combined	7020	1877	4513	2507

Table 3: Number of records per class in the four datasets used

The same schema was used to label the books in ECCO-classics, and the most frequently attested language-combinations are shown in Table 1 and 2. An example of multilingual work is for instance 'Phædri Augusti liberti Fabularum æsopiarum libri quinque. Or, a correct latin edition of the Fables of Phædrus: with a new literal English translation, and a copious parsing-index; Whereby young Beginners may easily and speedily attain the Knowledge of the Latin Tongue. By a gentleman of the University of Cambridge. For the Use of Schools', while an example of monolingual translation is given by 'The iliad of Homer. Translated by Alexander Pope, Esq.'.

4. Methodology

4.1. Tasks

As mentioned above, we aim at classifying the titles following two criteria, namely whether the edition is monolingual or multilingual (multilingual task henceforth), and whether, in case it is monolingual, it contains a work in its original language or in translation (monolingual translation task henceforth). We work with four combinations of the datasets, as listed in Table 3: the CAA, ECCO-classics, balanced CAA,[9] and ECCO and CAA combined. The datasets were split in 80-20 for training and test.

Multilingual and translated works are proportionately more frequent in the ECCO-classics dataset, because printing multilingual editions (i.e. the original text + a commentary or a translation in a modern language) was common practice for the circulation of classical works. When testing the different models, we evaluate the option of training on each dataset separately and

[9]We kept double the number of monolingual editions compared to monolingual translations in order to still achieve enough critical mass in the number of examples.

testing on each dataset separately, or training with the union of the two and testing on the datasets separately and combining them. In this way, we want to assess both the capacity of the separate models to generalize, and whether more increasing and diversifying the training data improved the final results (RQ3).

4.2. Models and approaches

In order to answer RQ 2, we have tested three different approaches: (1) a simple tf-idf model with Linear Support Vector classification [35] (ML henceforth), (2) fine-tuning a Large Language Model (BERT henceforth), and (3) taking a few-shot approach to fine-tune a sentence transformer model (SetFit henceforth). For the ML task, we performed minimal preprocessing of the titles (they were made lowercase, and punctuation was stripped), and created a common vocabulary comprising CAA and ECCO titles. We performed hyperparameter optimization for each model trained, on the hyperparameters ngram range (all combinations of monograms, bigrams and trigrams), the norm used for penalizing the model and avoiding overfitting ('l1', 'l2', 'elasticnet', None) and whether to weight the classes to limit the impact of very frequent classes ('weighted', None).

For the BERT approach, we fine-tuned the base model bert-base-multilingual-cased [7], using the HuggingFace API and packages. We used the model hyperparameters set out in the HuggingFace documentation for fine-tuning BERT for text classification [32], and for this paper, we have not performed hyperparameter optimization on them.

For the few-shot experiment the aim was to provide a small number of examples which were as representative as possible with respect to each task. Separate sets were made for the multilingual and translation tasks. For the multilingual task, the final training set contains 5 examples from each of the languages or language pairs, and an equal number of monolingual and multilingual titles, from both the ECCO and CAA datasets, resulting in about 80 examples in the train set. The train set for the translation task was constructed in a similar way but with an even number of original language and translated works. These were then evaluated using the same test sets as above.

To perform the few-shot classification, the SetFit library was used. SetFit fine-tunes a pre-trained SentenceTransformers model [29] using a contrastive training approach. SentenceTransformers is a form of Transformer-based Large Language Model which can be trained to generate embedding representations at the sentence, paragraph, or document level (rather than at the word-level as a regular LLM). These embeddings are then generally used for tasks such as semantic textual similarity or semantic search. SetFit is a framework for few-shot fine-tuning SentenceTransformers models. Setfit has shown to have performance comparable to a LLM-based approach on tasks such as text classification, but with far fewer data and training time [36]. We used the pre-trained SentenceTransformers model distiluse-base-multilingual-cased-v2 and the hyperparameters from the examples set out in the introductory guide [34]. We then fine-tuned the SentenceTransformers model using a small number of examples.

For each set of results we recorded the accuracy, as well as the precision, recall and f1 scores separately for each class. We include tables comparing the results of the two main tasks, plus the full tables as an appendix. Moreover, we used the SHAP (SHapley Additive exPlanations) library [21] to understand the features most relevant in the classification by the model. SHAP

is based on Shapely values, a game-theory approach to explanations which aims to calculate the contribution of each feature in an instance of a prediction. We used the SHAP library to produce plots which highlight tokens and spans of text based on their contribution to the prediction (Figure 1). These plots can then be interpreted qualitatively.

5. Results

5.1. Quantitative results

Below are shown some of the most relevant results, for the full set see the Appendix. Table 4 summarises the performance of the models trained on the 'combined' dataset and tested on both the individual and combined datasets. We report on the class-wise f-scores because the classes are very unevenly distributed, particularly for the CAA, and so the accuracy score is not a good indication of performance. Tables 5 and 6 give direct comparisons between the models on the multilingual and translation tasks, listing a difference simply by subtracting the score of the BERT model from the ML model (negative numbers mean the BERT model performed worse). Tables 7 to 10 in the Appendix provide the details of precision, recall and f1 for the ML and Bert models, on each task, for each class.

5.2. RQ1: Titles can be exploited for tracking multilingualism

As can be seen from Table 4, both BERT and the ML method gave quite comparable results across both tasks and all datasets. The SetFit method performed noticeably worse in most cases, except when tested on the combined CAA and ECCO dataset. Overall, results can be considered satisfactory which leads to the conclusion that titles can be used to this scope (RQ1), however the task requires an extended set of labeled training data to be provided.

5.3. RQ2: Comparison of the approaches

Tables 5 and 6 give direct comparisons between the models, listing a difference simply by subtracting the score of the BERT model from the ML model (negative numbers mean the BERT model performed worse). These show that generally, the tf-idf approach performed significantly better on the task to distinguish multilingual from monolingual works in many cases (with the exception of the set trained on the CAA and tested on ECCO). For the BERT model, in particular, Table 7 (in Appendix A.1) shows that the identification of the 0 class (i.e. multilingual works) is particularly problematic: recall values tend to be rather low - which indicates that the models tends to generally predict 'monolingual' for most titles.

For the translation task, there is slightly more variation between results of the approaches. The ML model has very low recall of the 1 class (translated work) when trained on the CAA and tested on ECCO, meaning almost all true positives (translations) are missed. This is a significant drawback since it is, for multilingualism studies, the class of interest. The BERT model performs reasonably well except again struggling with the recall of translated works when trained on the CAA and tested on another dataset. Most notable was the ability to identify ECCO translated documents using the model trained only on the CAA, both the full test dataset

and the smaller 'balanced' set, as well as the other way around. For this task, BERT was able to generalize much better than the ML method when testing on a different dataset than the one on which it was trained.

The performance of the setfit method (see the Appendix, Table 11) had a comparable pattern to the BERT models. It similarly had low recall and precision for the 0 class (multilingual works), but performed well with most tests on the translated works task, with just 40 examples of each class, across multiple languages.

5.4. RQ3: Specificity/generality of the training

In general, the ML and Bert models, when trained on examples from *across* datasets, are able to perform reasonably well - meaning that a training set made from a combined dataset of ECCO and the CAA gives satisfactory results. Both the ML method and the BERT fine-tuned model give very similar results.

Both models perform very well at identifying monolingual/multilingual works when trained and tested on ECCO. Models trained and tested on ECCO fared better in general, while still underperforming when applied to the CAA test dataset.

The results from models trained on one dataset and tested on the other are much worse. In particular, models trained on the CAA and tested on ECCO perform very badly at both recall and precision of the multilingual class. Again, there is little difference between the ML and BERT models, though the BERT model performs marginally better. The 'CAA balanced' model, trained on a sample of the CAA containing an equal number of monolingual/multilingual titles, balanced across the various target languages, did not perform significantly better than the CAA model, though it was marginally better and much quicker to train. However, the very small number of records might represent a limitation.

Since for the Setfit method we used a mix of examples coming from both datasets, RQ3 does not apply to this model.

5.5. Qualitative results

To understand qualitatively what parts of the text caused the classification, we use SHAP explanations, and looked at a range of true positive, true negative, false positive and false negative predictions. Here, we focus on the BERT model trained on the CAA and tested on both CAA and ECCO for the prediction of multilingual texts (a particularly 'difficult' combination).

When the models wrongly label a title as monolingual when it is multilingual, in general, these phenomena seem to occur:

- There is no trace of multilingualism in the title (e.g. the Latin title 'Specimen doctrine traditae ab anno MDCXCI.usque ad annum MDCXCVI. inclusive.' doesn't contain any mention of parts in a different language).
- Most of these titles, despite containing hints of multilingualism, are fully in Latin. The wrong prediction might be due to the fact that the CAA contains a lot of Latin monolingual titles, and hence Latin context is considered monolingual despite possible multilingual records. Figure 2 shows a very long title in Latin with an explicit mention of a

		Multilingual Task		Translation Task	
Train	Test	F-score (0)	F-score (1)	F-score (0)	F-score (1)
ML					
combined	caa	0.82	0.99	0.99	0.89
combined	ecco	0.91	0.97	0.98	0.99
combined	combined	0.75	0.97	0.98	0.96
BERT					
combined	caa	0.81	0.99	1.00	0.99
combined	ecco	0.91	0.97	0.99	0.90
combined	combined	0.78	0.97	0.98	0.96
SetFit					
Few-shot	caa	0.16	0.90	0.98	0.06
Few-shot	ecco	0.51	0.74	0.59	0.33
Few-shot	combined	0.42	0.82	0.80	0.23

Table 4: Class-wise f-scores for the fine-tuned BERT, SVM, and SetFit methods using combined CAA + ECCO datasets.

Train	Test	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f1 (1)
caa	caa	0.00	0.06	0.01	0.04	0.00	0.00	0.41
caa	ecco	0.00	0.18	-0.52	0.25	-0.06	0.03	-0.01
caa	combined	-0.01	0.09	-0.19	0.06	-0.01	0.01	0.00
caa	caa_balanced	-0.14	-0.34	-0.10	-0.24	-0.04	-0.14	-0.10
ecco	ecco	0.03	0.03	0.06	0.05	0.02	0.01	0.01
ecco	caa	-0.18	0.56	0.02	0.11	-0.22	0.02	-0.12
ecco	combined	-0.14	-0.10	-0.50	-0.37	-0.15	-0.01	-0.09
ecco	caa_balanced	0.07	0.34	-0.62	0.24	-0.38	0.22	0.02
combined	combined	0.00	0.08	-0.03	0.03	-0.01	0.01	0.00
combined	caa	0.01	0.06	-0.13	-0.01	-0.01	0.00	0.00
combined	ecco	0.00	0.00	0.01	0.00	0.00	0.00	0.00
combined	caa_balanced	0.08	0.20	-0.13	0.05	-0.08	0.22	0.09
caa_balanced	caa_balanced	-0.10	-0.09	-0.28	-0.17	-0.10	0.09	0.00
caa_balanced	caa	-0.39	-0.09	-0.26	-0.36	-0.39	-0.01	-0.27
caa_balanced	ecco	0.01	0.06	0.03	0.05	0.00	0.02	0.00

Table 5: Comparative results for the monolingual/multilingual task, for bert-base-multilingual-cased approach and tf-idf/SVM. Number reported is the BERT result subtracted from the tf-idf result. Numbers under zero mean that the BERT approach performed worse. Acc, r, p, and f1 denote accuracy, recall, precision, and f-score respectively.

Train	Test	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f1 (1)
caa	caa	0.02	0.01	0.01	0.01	0.22	0.17	0.20
caa	ecco	0.38	-0.04	0.22	0.19	0.60	-0.03	0.73
caa	combined	0.21	0.00	0.17	0.11	0.58	0.12	0.65
caa	caa_balanced	0.03	0.01	0.03	0.02	0.06	0.03	0.05
ecco	ecco	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00
ecco	caa	0.19	0.21	0.00	0.17	-0.07	0.03	0.04
ecco	combined	0.11	0.17	-0.01	0.12	0.00	0.12	0.08
ecco	caa_balanced	0.07	0.09	0.02	0.09	0.00	0.04	0.03
combined	combined	0.00	-0.01	0.01	0.00	0.01	-0.02	0.00
combined	caa	0.00	0.01	0.00	0.01	0.00	0.02	0.01
combined	ecco	0.01	0.03	-0.01	0.01	-0.01	0.02	0.00
combined	caa_balanced	0.02	0.01	0.02	0.01	0.03	0.03	0.04
caa_balanced	caa_balanced	0.01	-0.04	0.05	0.00	0.12	-0.04	0.04
caa_balanced	caa	-0.01	-0.02	0.00	-0.01	0.00	-0.07	-0.07
caa_balanced	ecco	0.48	-0.13	0.45	0.30	0.85	-0.08	0.82

Table 6: Comparative results for the translation task, for bert-base-multilingual-cased approach and tf-idf/SVM. Number reported is the BERT result subtracted from the tf-idf result. Numbers under zero mean that the BERT approach performed worse. Acc, r, p, and f1 denote accuracy, recall, precision, and f-score respectively.

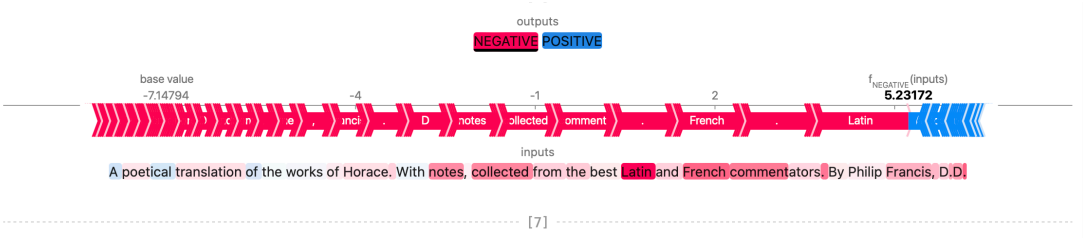

Figure 1: Example of a text plot from the python SHAP library. In this case, parts of the text contributing to the identification of the title as a translation are highlighted in red.

translated bit ('cum latina interpretatione') being entirely assigned to monolingual (blue) by the model.

Another recurrent trend in both false and true prediction is the role of Greek: the word 'Greek' (or *Gracae*, *in Graecam linguam*) is always used as a predictor of multilingualism, even when the work is monolingual (either in the original language or in translation). Figure 4 and 3 show an example of two monolingual works whose titles contain the word 'Greek'. In both cases, the word Greek heavily impacts the 'multilingual' component, despite the fact that the output is different for the two predictions. This might be due to the fact that in the CAA Ancient Greek texts usually come with translations/notes in a modern language. Text in the Greek alphabet also seems to be used to make identifications of multilingual texts. This raises

Figure 2: Example of a text plot from the python SHAP library. In this case, parts of the text contributing to the identification of the title as multilingual are highlighted in red. The title was labeled as monolingual while being multilingual.

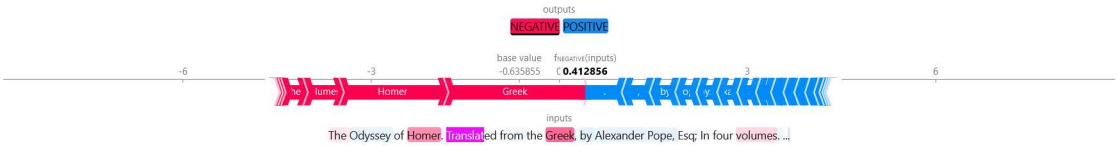

Figure 3: Example of a text plot from the python SHAP library. In this case, parts of the text contributing to the identification of the title as multilingual are highlighted in red. The title was labeled as multilingual while being monolingual. The word Greek heavily contributes to the multilingual prediction

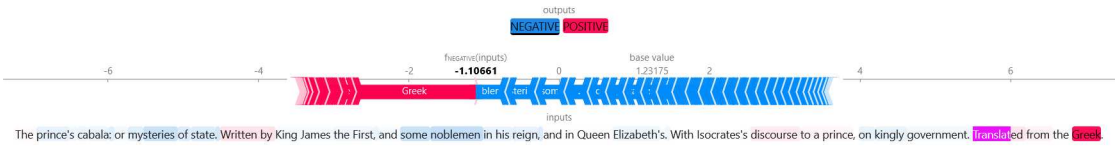

Figure 4: Example of a text plot from the python SHAP library. In this case, parts of the text contributing to the identification of the title as multilingual are highlighted in red. The title was correctly labeled as monolingual, but the word Greek heavily contributes to the multilingual prediction

the issue of the dependency of the models on these specific dataset features. Furthermore, the model in some cases uses the text which we would read as making it likely to be multilingual as an output pointing to monolingual. For example things like 'original subjoined' or 'notes at the end', 'on the opposite page'... One example of this can be seen in Figure 5. This is because these phrases are not found in the CAA titles for multilingual works. The 'combined' model doesn't have this bias, in this case, words relating to notes or annotations contribute to a positive prediction of a work as multilingual, as one might expect.

Words like 'translated', or 'lexicon' across languages increase the output of the model in identifying multilingual works, which is close to what we would expect.

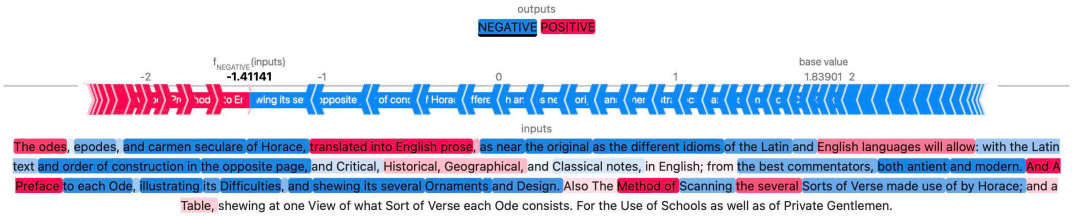

Figure 5: Example of SHAP plot showing a work from ECCO predicted as monolingual by the CAA-trained model. Parts of the text which we would intuitively see make it likely to be multilingual are in fact in this cases contributing to the prediction of the instance as monolingual.

6. Discussion of relevance and possible uses

Overall, these experiments suggest it is a difficult problem to solve using machine learning methods. In particular, the approaches do not seem to generalise well, even using multilingual LLMs which we hoped might mean that different styles of title would be recognised if they were in some way semantically similar. This is perhaps because the way that multilingual and translated works are signified in a title is varied and changes over time and across languages. Despite these reservations, when trained on examples across both datasets, the performance of both traditional machine learning and LLM methods was at a level which we deem usable in real-world applications.

The multilingual fine-tuned BERT has some advantages over traditional ML approaches in identifying translated works but performs worse when distinguishing multilingual works. This seems to be because the signifiers for translated works are more descriptive and straightforward (e.g. 'translated from' or 'made English by'). The multilingual approach means that these kinds of phrases tend be be picked up by the model in different languages.

The few-shot method using SetFit shows some promise in a number of tasks, but does not, from our experiments, seem to be a 'silver bullet' for low-resource metadata enrichment of this kind. However, perhaps with a very well thought-out and diverse set of examples, it may be possible to build a model which can be trained and used for inferences on real-world data. An ideal real-world scenario for metadata enrichment may involve collecting a small number of examples from a specific dataset or collection, fine-tuning a bespoke but small model, and applying it only to that collection. However, as of yet, from our experiments, it does not seem that the multilingual capabilities of SetFit or SentenceTransformers are enough to get high-quality results on this task without at least some annotation of the target dataset.

7. Conclusions

Automatically enriched metadata has significant value to heritage collections catalogue data, potentially helping to increase the accuracy and findability of records. If the purpose is to get enriched metadata, our experiments show some promise and could potentially be operationalised in the future. In fact, traditional ML methods may be enough in many cases, partic-

ularly for identifying multilingual works, and have big advantages in terms of ease of use and use of resources. In some cases, methods such as keyword search or regular expressions might also provide acceptable results, though when using multilingual datasets, machine learning methods should have an advantage.

Furthermore, we suggest that certain evaluation metrics are more important than others, particularly with library catalogue data, which is likely to be very unevenly distributed with regards to language and classes. This is of course dependant on the particular task and use-case. If the purpose is to improve catalogue metadata for example, the recall of the multilingual or translated classes may be particularly important, as it may be better to find additional false positives which can then be checked manually afterwards, rather than aiming for precision but missing some relevant works. If the information is not necessarily intended to be 'fed back' to a catalogue but used for bibliographic data science at scale, it may be more important to focus on the overall f-scores to get a broad, albeit imperfect, accuracy.

Acknowledgments

We want to express our gratitute to the STUDIUM.AI team, particular to Violet Soen, whose efforts enabled this research. In addition, we would like to thank the KU Leuven Libraries staff, in particular the metadata and digitization services for sharing the CAA metadata and the relative documentation. Finally, we would like to thank the Computational History group of Helsinki, for providing the framework and infrastructure for annotating the ECCO training data.

References

[1] D. Ali, K. Milleville, S. Verstockt, N. Van De Weghe, S. Chambers, and J. M. Birkholz. "Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections". In: *Journal of Documentation* (2023). DOI: 10.1108/jd-01-2022-0029.

[2] P. Auger and S. Brammall, eds. *Multilingual texts and practices in early modern Europe*. New York, NY: Routledge, 2023.

[3] T. W. Baldwin. *William Shakspere's Small Latine and Lesse Greeke*. Urbana: University of Illinois Press, 1944.

[4] B. Bistué. "Collaborative Translation as a Model for Multilingual Printing in Early Renaissance Editions of Aesop's Fables". In: *Multilingual texts and practices in early modern Europe*. Ed. by P. Auger and S. Brammall. New York, NY: Routledge, 2023.

[5] M. Bowman. "Text-mining metadata: What can titles tell us of the history of modern and contemporary art?" In: *Journal of Cultural Analytics* 8.1 (2023). DOI: 10.22148/001c.74602.

[6] N. Constantinidou. "Printers of the Greek Classics and Market Distribution in the Sixteenth Century: The Case of France and the Low Countries". In: *Specialist Markets in the Early Modern Book World* 40 (2015). Ed. by R. Kirwan and S. Mullins, pp. 273–93.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[9] M. Fantoli, J. Suomela, T. Van Hal, M. Depauw, L. Virkki, and M. Tolonen. “Quantifying the Presence of Ancient Greek and Latin Classics in Early Modern Britain”. In: *Journal of Cultural Analytics* (forthcoming).

[10] C. Garcia-Zorita and A. R. Pacios. “Topic modelling characterization of Mudejar art based on document titles”. In: *Digital Scholarship in the Humanities* 33.3 (2018), pp. 529–539. DOI: 10.1093/llc/fqx055.

[11] S. Gillespie. “The Availability of the Classics. Readers, Writers, Translation, Performance”. In: *The Oxford History of Classical Reception in English Literature. 1558-1660*. Vol. 2. Oxford University Press, 2015, pp. 57–74.

[12] H. O. Hatzel, H. Stiemer, C. Biemann, and E. Gius. “Machine learning in computational literary studies”. In: *it - Information Technology* 65.4-5 (2023), pp. 200–217. DOI: 10.1515/itit-2023-0041.

[13] T. Hermans. “Multilingualism and Translation in the Early Modern Low Countries”. In: *Language Dynamics in the Early Modern Period*. Ed. by K. Bennett and A. Cattaneo. 1st ed. New York: Routledge, 2022, p. 20. DOI: 10.4324/9781003092445. URL: https://www.taylorfrancis.com/books/9781003092445.

[14] R. Hoven. “Enseignement du grec et livres scolaires dans les anciens Pays-Bas et la Principaute de Liege de 1483 à 1600. Deuxième partie: 1551-1600”. In: *Gutenberg-Jahrbuch* 55 (1980), pp. 118–26.

[15] R. Hoven. “Enseignement du grec et livres scolaires dans les anciens Pays-Bas et la Principauté de Liège de 1483 à 1600. Première partie: 1483-1550”. In: *Gutenberg-Jahrbuch* 54 (1979), pp. 80–86.

[16] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, and K. Lindén. “Automatic Language Identification in Texts: A Survey”. In: *Journal of Artificial Intelligence Research* 65 (2019). DOI: 10.1613/jair.1.11675.

[17] H. Jones. “Printing the Classical Text”. In: *Printing the Classical Text*. Brill, 2021. URL: https://brill.com/display/title/26045.

[18] L. Lahti, N. Ilomäki, and M. Tolonen. “A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800”. In: *LIBER Quarterly: The Journal of the Association of European Research Libraries* 25.2 (2015), pp. 87–116. DOI: 10.18352/lq.10112.

[19] L. Lahti, J. Marjanen, H. Roivainen, and M. Tolonen. "Bibliographic Data Science and the History of the Book (c. 1500–1800)". In: *Cataloging & Classification Quarterly* 57.1 (2019), pp. 5–23. DOI: 10.1080/01639374.2018.1543747.

[20] H. B. Lathrop. *Translations from the Classics into English from Caxton to Chapman (1477-1620)*. Vol. 35. University of Wisconsin Studies in Language and Literature. Madison: University of Wisconsin, 1933.

[21] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[22] P. Mack. "Humanism and the Classical Tradition". In: *The Oxford History of the Renaissance*. Ed. by G. Campbell. 1st ed. Oxford University PressOxford, 2023, pp. 10–47. DOI: 10.1093/oso/9780192886699.003.0001.

[23] V. Malínek, T. Umerle, E. Gray, I. Heibi, P. Király, C. Klaes, P. Korytkowski, D. Lindemann, A. Moretti, C. Panušková, R. Péter, M. Tolonen, A. Tomczyńska, and O. Vimr. "Open Bibliographical Data Workflows and the Multilinguality Challenge". In: *Journal of Open Humanities Data* 10 (2024), p. 27. DOI: 10.5334/johd.190.

[24] M. Martorana, T. Kuhn, L. Stork, and J. van Ossenbruggen. *Text classification of column headers with a controlled vocabulary: leveraging LLMs for metadata enrichment*. 2024. URL: http://arxiv.org/abs/2403.00884.

[25] F. Moretti. "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740?1850)". In: *Critical Inquiry* 36.1 (2009), pp. 134–158. DOI: 10.1086/606125.

[26] J. A. Nolazco-Flores, A. V. Guerrero-Galván, C. Del-Valle-Soto, and L. P. Garcia-Perera. "Genre Classification of Books on Spanish". In: *IEEE Access* 11 (2023), pp. 132878–132892. DOI: 10.1109/access.2023.3332997.

[27] R. Péter, Z. Szántó, Z. Biacsi, G. Berend, and V. Bilicki. "Multilingual Analysis and Visualization of Bibliographic Metadata and Texts With the AVOBMAT Research Tool". In: *Journal of Open Humanities Data* 10 (2024), p. 23. DOI: 10.5334/johd.175.

[28] I. Rastas, Y. Ciarán Ryan, I. Tiihonen, M. Qaraei, L. Repo, R. Babbar, E. Mäkelä, M. Tolonen, and F. Ginter. "Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model". In: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Ed. by N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, and L. Borin. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 68–77. DOI: 10.18653/v1/2022.lchange-1.7.

[29] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[30] Y. C. Ryan and M. Tolonen. "The Evolution of Scottish Enlightenment Publishing". In: *The Historical Journal* 67.2 (2024), pp. 223–255. DOI: 10.1017/s0018246x23000614.

[31] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu. *Large Language Models for Data Annotation: A Survey*. 2024. DOI: 10.48550/arxiv.2402.13446.

[32] *Text classification*. 2024. URL: https://huggingface.co/docs/transformers/en/tasks/sequence%5C%5Fclassification.

[33] M. Tolonen, E. Mäkelä, and L. Lahti. "The Anatomy of Eighteenth Century Collections Online (ECCO)". In: *Eighteenth-Century Studies* 56.1 (2022), pp. 95–123. DOI: 10.1353/ecs.2022.0060.

[34] L. Tunstall. *SetFit: Efficient Few-Shot Learning Without Prompts*. 2022. URL: https://huggingface.co/blog/setfit.

[35] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg. "Efficient Few-Shot Learning Without Prompts". In: (2022). DOI: 10.48550/arxiv.2209.11055.

[36] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg. *Efficient Few-Shot Learning Without Prompts*. 2022. DOI: 10.48550/arxiv.2209.11055.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is All You Need". In: 2017. URL: https://arxiv.org/pdf/1706.03762.pdf.

[38] *Vertalen in de Nederlanden: een cultuurgeschiedenis*. Amsterdam: Boom, 2021.

[39] F. Watson. *The English Grammar Schools to 1660. Their Curriculum and Practice*. 2nd ed. London: Frank Cass & Co., 1968.

[40] M. Wevers, N. Vriend, and A. De Bruin. "What to do with 2.000.000 Historical Press Photos? The Challenges and Opportunities of Applying a Scene Detection Algorithm to a Digitised Press Photo Collection". In: *TMG Journal for Media History* 25.1 (2022), p. 1. DOI: 10.18146/tmg.815.

[41] P. Wilson. "The Place of Classics in Education and Publishing". In: *The Oxford History of Classical Reception in English Literature. 1660-1790*. Ed. by D. Hopkins and C. Martindale. Vol. 3. Oxford and New York: Oxford University Press, 2012, pp. 29–52.

[42] J. Zhang, Y. C. Ryan, I. Rastas, F. Ginter, M. Tolonen, and R. Babbar. "Detecting Sequential Genre Change in Eighteenth-Century Texts". In: *Proceedings of the Computational Humanities Research Conference 2022*. Ed. by F. Karsdorp, A. Lassche, and K. Nielbo. Vol. 3290. CEUR Workshop Proceedings. Antwerp, Belgium: Ceur, 2022, pp. 243–255. URL: https://ceur-ws.org/Vol-3290/%5C#short%5C%5Fpaper2630.

8. Appendix

A. Full Results

A.1. Multilingual/Monolingual Task: BERT

Train	Test	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f1 (1)
caa	caa	0.96	0.62	0.59	0.61	0.98	0.98	0.98
caa	ecco	0.77	0.19	0.48	0.27	0.94	0.80	0.86
caa	combined	0.89	0.35	0.69	0.46	0.98	0.91	0.94
caa	caa_balanced	0.85	0.64	0.90	0.75	0.96	0.83	0.89
ecco	ecco	0.93	0.79	0.88	0.84	0.97	0.94	0.95
ecco	caa	0.75	0.62	0.10	0.18	0.75	0.98	0.85
ecco	combined	0.82	0.81	0.42	0.55	0.83	0.97	0.89
ecco	caa_balanced	0.55	0.43	0.38	0.40	0.62	0.67	0.64
combined	combined	0.94	0.74	0.82	0.78	0.97	0.96	0.97
combined	caa	0.98	0.78	0.83	0.81	0.99	0.99	0.99
combined	ecco	0.96	0.90	0.93	0.91	0.98	0.97	0.97
combined	caa_balanced	0.93	0.93	0.87	0.90	0.92	0.96	0.94
caa_balanced	caa_balanced	0.75	0.64	0.64	0.64	0.81	0.81	0.81
caa_balanced	caa	0.53	0.88	0.08	0.14	0.52	0.99	0.68
caa_balanced	ecco	0.62	0.45	0.29	0.36	0.68	0.81	0.73
caa_balanced	combined	0.58	0.55	0.17	0.26	0.59	0.89	0.71

Table 7: Performance results for Monolingual/Multilingual task and fine-tuned BERT

A.2. Multilingual/Monolingual Task: TFIDF/SVM

Train	Test	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f1 (1)
caa	caa	0.96	0.56	0.58	0.57	0.98	0.98	0.57
caa	ecco	0.77	0.01	1.00	0.02	1.00	0.77	0.87
caa	combined	0.90	0.26	0.88	0.40	0.99	0.90	0.94
caa	caa_balanced	0.99	0.98	1.00	0.99	1.00	0.97	0.99
ecco	ecco	0.90	0.76	0.82	0.79	0.95	0.93	0.94
ecco	caa	0.93	0.06	0.08	0.07	0.97	0.96	0.97
ecco	combined	0.96	0.91	0.92	0.92	0.98	0.98	0.98
ecco	caa_balanced	0.48	0.09	1.00	0.16	1.00	0.45	0.62
combined	combined	0.94	0.66	0.85	0.75	0.98	0.95	0.97
combined	caa	0.97	0.72	0.96	0.82	1.00	0.99	0.99
combined	ecco	0.96	0.90	0.92	0.91	0.98	0.97	0.97
combined	caa_balanced	0.85	0.73	1.00	0.85	1.00	0.74	0.85
caa_balanced	caa_balanced	0.85	0.73	0.92	0.81	0.91	0.72	0.81
caa_balanced	caa	0.92	0.97	0.34	0.50	0.91	1.00	0.95
caa_balanced	ecco	0.61	0.39	0.26	0.31	0.68	0.79	0.73

Table 8: Performance results for multilingual/monolingual task and TFIDF/SVM

A.3. Translation Task: BERT

Train	Test	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f1 (1)
caa	caa	0.98	0.99	0.99	0.99	0.74	0.72	0.73
caa	ecco	0.75	0.96	0.59	0.73	0.62	0.97	0.76
caa	combined	0.87	0.99	0.83	0.90	0.64	0.98	0.77
caa	caa_balanced	0.97	1.00	0.96	0.98	0.91	1.00	0.95
ecco	ecco	0.96	0.91	0.97	0.94	0.99	0.95	0.97
ecco	caa	0.66	0.65	0.99	0.78	0.87	0.10	0.18
ecco	combined	0.83	0.73	0.99	0.84	0.99	0.68	0.80
ecco	caa_balanced	0.61	0.47	0.92	0.62	0.91	0.44	0.59
combined	combined	0.97	0.97	0.99	0.98	0.97	0.95	0.96
combined	caa	0.99	1.00	1.00	1.00	0.90	0.90	0.90
combined	ecco	0.99	0.99	0.98	0.99	0.99	1.00	0.99
combined	caa_balanced	0.96	1.00	0.95	0.97	0.88	1.00	0.94
caa_balanced	caa_balanced	0.87	0.86	0.94	0.90	0.88	0.74	0.81
caa_balanced	caa	0.93	0.92	1.00	0.96	0.97	0.37	0.54
caa_balanced	ecco	0.88	0.87	0.82	0.84	0.89	0.92	0.91
caa_balanced	combined	0.91	0.90	0.96	0.93	0.93	0.85	0.89

Table 9: Performance results for translation task and fine-tuned BERT

A.4. Translation Task: TFIDF/SVM

Train	Test	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f-score (1)
caa	caa	0.96	0.98	0.98	0.98	0.52	0.55	0.53
caa	ecco	0.37	1.00	0.37	0.54	0.02	1.00	0.03
caa	combined	0.66	0.99	0.66	0.79	0.06	0.86	0.12
caa	caa_balanced	0.94	0.99	0.93	0.96	0.85	0.97	0.90
ecco	ecco	0.96	0.91	0.98	0.95	0.99	0.95	0.97
ecco	caa	0.47	0.44	0.99	0.61	0.94	0.07	0.14
ecco	combined	0.72	0.56	1.00	0.72	0.99	0.56	0.72
ecco	caa_balanced	0.54	0.38	0.90	0.53	0.91	0.40	0.56
combined	combined	0.97	0.98	0.98	0.98	0.96	0.97	0.96
combined	caa	0.99	0.99	1.00	0.99	0.90	0.88	0.89
combined	ecco	0.98	0.96	0.99	0.98	1.00	0.98	0.99
combined	caa_balanced	0.94	0.99	0.93	0.96	0.85	0.97	0.90
caa_balanced	caa_balanced	0.86	0.90	0.89	0.90	0.76	0.78	0.77
caa_balanced	caa	0.94	0.94	1.00	0.97	0.97	0.44	0.61
caa_balanced	ecco	0.40	1.00	0.37	0.54	0.04	1.00	0.09

Table 10: Performance results for translation task and TFIDF/SVM

A.5. SetFit model results

Test set	Acc	r (0)	p (0)	f1 (0)	r (1)	p (1)	f1 (1)
Monolingual/Multilingual Task							
caa	0.49	0.41	1.00	0.59	1.00	0.20	0.33
ecco	0.96	0.96	1.00	0.98	1.00	0.03	0.06
combined	0.68	0.67	1.00	0.80	0.96	0.13	0.23
Translation Task							
caa	0.82	0.10	0.41	0.16	0.97	0.84	0.90
ecco	0.66	0.38	0.78	0.51	0.91	0.62	0.74
combined	0.73	0.30	0.74	0.42	0.95	0.73	0.82

Table 11: Performance results for SetFit model, trained on a small diverse sample and tested on CAA/ECCO/Combined datasets.