Classifying Textual Genre in Historical Magazines (1875-1990)

Vera Danilova and **Ylva Söderfeldt**
Uppsala University, Department of History of Science and Ideas, Uppsala, Sweden
{vera.danilova, ylva.soderfeldt}@idehist.uu.se

Abstract

Historical magazines are a valuable resource for understanding the past, offering insights into everyday life, culture, and evolving social attitudes. They often feature diverse layouts and genres. Short stories, guides, announcements, and promotions can all appear side by side on the same page. Without grouping these documents by genre, term counts and topic models may lead to incorrect interpretations. This study takes a step towards addressing this issue by focusing on genre classification within a digitized collection of European medical magazines in Swedish and German. We explore two scenarios: 1) leveraging the available web genre datasets for zero-shot genre prediction, 2) semi-supervised learning over the few-shot setup. This paper offers the first experimental insights in this direction. We find that 1) with a custom genre scheme tailored to historical dataset characteristics it is possible to effectively utilize categories from web genre datasets for cross-domain and cross-lingual zero-shot prediction, 2) semi-supervised training gives considerable advantages over few-shot for all models, particularly for the historical multilingual BERT. The models and code are available on GitHub[1].

1 Introduction

Quantitative processing of digitized archives referred to as "distant reading" helps historians in conducting large-scale analysis and categorization of their data (Moretti, 2000). However, it is of utmost importance to develop methods that can contribute to reliable interpretations (Da, 2019). This paper proposes genre[2] classification to improve reliability of distant reading interpretations for visually- and information-rich materials, such as historical magazines.

The ActDisease[3] Dataset is an extensive private collection currently being digitized as part of an ongoing project on the modern history of European medicine. It consists of medical magazines issued by ten European patient organizations in four languages (Swedish, German, French, and English) throughout the 20th century. Each magazine had a different publication frequency, resulting in a varying number of issues per year and featuring diverse page formats and visually complex layouts. Within the same page, these magazines often combine texts that carry different communicative purposes, such as personal narratives, advertisements, instructions, short stories, etc. Failing to group these texts hinders the accurate interpretation of term counts and topic models across historical periods, as results may be skewed toward the most frequent genres.

Moreover, grouping by genre enriches historical interpretations by providing a broader view of evolving communicative strategies over time (Broersma, 2010) and allowing for more fine-grained analyses of term distributions and topic models. In the historical research, these text groups are referred to as "epistemic genres" and have been recognized as a valuable conceptual tool for exploring cross-cultural history of medicine (Pomata, 2014; Hanson, 2022). Their use is linked to the development of knowledge communities, such as patient organizations in our case.

Due to scarcity of annotated data for our dataset, this paper explores the effectiveness of zero-shot learning using publicly available datasets annotated with web genres and registers, and assesses the mapping of the existing categories to our custom ones. We also investigate the impact of few-shot learning, comparing the standard approach with a semi-supervised method that incorporates prior

[1] https://github.com/veraDanilova/genre-classification-LaTeCH-CLFL2025

[2] We use the definition of genre as a class of documents that share a communicative purpose (Kessler et al., 1997)

[3] ActDisease project (ERC-2021-STG 10104099): http://actdisease.org

Proceedings of LaTeCH-CLfL 2025, pages 160–171
May 4, 2025 ©2025 Association for Computational Linguistics

fine-tuning on the full dataset to leverage broad knowledge and task-specific adaptation. Additionally, we compare the performance of models pretrained on modern versus historical data in classifying genres within historical materials.

The paper is organized as follows. Section 2 discusses work on genre classification for historical materials, as well as the recent advances in genre classification with LLMs. Section 3 provides definitions for genre categories. Section 4 gives details on the datasets used for zero-shot and few-shot experiments. Section 5 outlines the experimental setup: models and types of experiments. Section 6 discusses the results, and Section 7 concludes the paper.

2 Related Work

The application of classical machine learning (ML) methods to genre extraction from historical newspapers has been discussed in (Broersma and Harbers, 2018). Rather than proposing a specific algorithmic solution, the authors focus on the challenges of defining and annotating genres in historical newspapers, as well as the difficulties in transparently evaluating and comparing different algorithms.

While automatic genre classification for historical sources remains relatively unexplored, there is extensive research on automatic identification of web genres and registers (Kuzman and Ljubešić, 2023). The state-of-the-art textual genre classifier is a version of XLM-Roberta (Conneau et al., 2020) fine-tuned on a combination of web genre datasets with extensive genre category coverage (Kuzman et al., 2023).

We extend this work by comparing the performance of three multilingual encoders - XLM-Roberta, mBERT (Devlin et al., 2019) and historical mBERT (Schweter et al., 2022) - for zero-shot and few-shot genre classification of multilingual historical magazines.

3 Genre Categories

A set of genre categories was defined under the supervision of the main historian of the project who specializes in patient organizations. *Academic* reports about academic research or explains complex scientific ideas in an accessible way (research article, report or popular science article). *Administrative* reports about the activities or operations of the patient organizations (meeting minutes, financial reports, annual reports, editorial information, official correspondence and petitions, announcements). *Advertisement* promotes products or services with intent to sell them (promotion, advertisement). *Guide* provides advice or instructions for step-by-step implementation to achieve a certain goal or solve a problem related to health, legal issues or other (dietary advice, physical exercise instructions, recipe, procedural instructions, application guidelines). *Fiction* aims to entertain the reader, gives reading pleasure, engages the reader emotionally (poems, short stories, humor, myths, novel, novellas). *Legal* explains or informs about terms and conditions (contract, rules, amendment, general terms and conditions). *News* informs or reports about updates on recent events and important developments (daily news). *Nonfiction prose* (nf_prose) narrates about events or experiences from personal life or represents a description of cultural phenomena or history (historical narrative, auto(biography), memoire, travel note, personal letter, opinion essay, cultural article, documentary prose). *QA* is text structured in a question-answer format, for example questions from members and answers from medical professionals.

4 Datasets

For zero-shot prediction, we take advantage of the publicly available datasets used in the previous work for automatic genre identification (AGI) (Kuzman et al., 2023) and the investigation of cross-lingual genre transfer in dependency parsing (Danilova and Stymne, 2023).

The entire ActDisease Dataset is used to fine-tune the models for masked language modeling (MLM) in the semi-supervised few-shot scenario. A portion of it is annotated for the experiments.

4.1 AGI Datasets

Corpus of Online Registers of English (CORE) (Egbert et al., 2015) is a large dataset containing around 50k documents manually annotated with web registers. It uses a two-level label hierarchy: 8 main registers and 47 subregisters. Subregisters are fine-grained and are well-suitable for mapping to our categories. The mapping is discussed in detail in the next subsection that describes the historical dataset. Multilingual register corpora in Swedish, Finnish and French (Repo et al., 2021) are annotated only with the main registers and we leverage them for mapping only partially.

Functional Text Dimensions (FTD) is a dataset

Historical	CORE	UDM	FTD
academic	research article (RA)	academic	academic (A14)
administrative	-	parliament	-
advertisement	advertisement (AD), description with intent to sale (DS)	-	commercial (A12)
guide	how-to (HT), recipe (RE), other how-to/instructional (OH), how-to instructional (HI)	guide	instruct (A7)
fiction	poem (PO), short story (SS)	fiction	fictive (A4), poetic (A19)
legal	legal terms and conditions (LT)	legal	legal (A9)
news	news report / blog (NE)	news	reporting (A8)
nonfiction_prose	personal blog (PB), opinion blog (OB), travel blog (TB), historical article (HA), magazine article (MA)	nonfiction prose, blog	personal (A11), argumentative (A1)
QA	question/answer forum (QA), advice (AV)	QA	

Table 1: Mapping of genre categories between the AGI datasets and the ActDisease Dataset

DATA	*G*	*B*	*instances*	*tokens*
CORE	+	2	28.5K	7.5M
CORE	+	1	33.7K	8.7M
CORE	-	1	33.6K	8.7M
CORE	-	2	25.8K	6.7M
FTD	+	2	3.8K	1.0M
FTD	-	1	7.0K	1.7M
FTD	+	1	3.8K	1.0M
FTD	-	2	7.0K	1.7M
UDM	-	1	5.0K	1.0M
UDM	+	1	1.4K	0.3M
UDM	+	2	1.3K	0.3M
UDM	-	2	5.0K	1.0M
merged	+	1	40.2K	10.4M
merged	+	2	24.2K	6.3M
merged	-	2	40.1K	9.7M
merged	-	1	55.6K	13.8M

Table 2: Training data configurations for the AGI datasets. `[B2]` means balanced by two levels: our label and original dataset labels. `[B1]` means balancing by our labels only. `[G+]` means the filtering by language family is performed and only Germanic languages are present in the dataset. `[G-]` is for the case when all language families are included in the dataset

DATA	*language family*	*instances*	*tokens*
CORE	gem	3.9K	1041.1K
FTD	gem, sla	700	174.8K
UDM	gem, roa, sla, urj	720	156.9K
merged	gem, roa, sla, urj	4.6K	1084.4K

Table 3: Test data for intra-dataset evaluation of the classifiers. In the language family column, we use ISO codes of languages families: gem - Germanic, roa - Romance, sla - Slavic, urj - Uralic

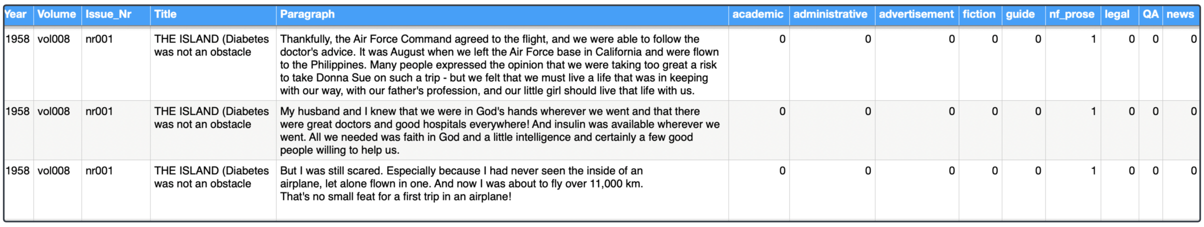
of document-level annotations of web genres (Sharoff, 2021; Lepekhin and Sharoff, 2022). We use the available data for two languages: English and Russian. Documents belonging to multiple labels or annotated as "unsuitable" are discarded. The final dataset includes 1686 English and 1693 Russian documents labeled with 10 categories.

UD-MULTIGENRE (UDM) is a subset of Universal Dependencies (UD) in 38 languages enriched with genre annotations on sentence level (Danilova and Stymne, 2023). It uses 17 genre categories based on the original treebank-level UD labels and contains 657.4k sentences (11M tokens) in total.

X-GENRE dataset is a combination of English CORE, English FTD and Slovene Ginco (Kuzman et al., 2023). Since there already exists a model for genre classification fine-tuned on this dataset, we use it as a baseline for zero-shot prediction on comparable categories.

Genre Category Mapping. To map the categories of the datasets to genres of our historical dataset, two annotators independently reviewed the guidelines of each dataset and assigned the most suitable categories to our genre labels. The categories on which the annotators agree are grouped under the corresponding genres. The final mapping is presented in Table 1.

Dataset sampling configurations. We investigate two aspects of the training data: language selection and data balancing. For language selection, we consider two scenarios: training on the entire set of available languages and training exclusively on Germanic languages. In terms of data balancing, we implement two strategies. The first involves balancing the data at two levels: our genre labels and the corresponding AGI labels. This ensures that all AGI subcategories are equally represented. The second strategy focuses on balancing only by our genre labels, using downsampling to reduce the size of the largest genres. The configurations are shown in Table 2. *Merged* signifies that, for



Year	Volume	Issue_Nr	Title	Paragraph	academic	administrative	advertisement	fiction	guide	nf_prose	legal	QA	news
1958	vol008	nr001	THE ISLAND (Diabetes was not an obstacle	Thankfully, the Air Force Command agreed to the flight, and we were able to follow the doctor's advice. It was August when we left the Air Force base in California and were flown to the Philippines. Many people expressed the opinion that we were taking too great a risk to take Donna Sue on such a trip - but we felt that we must live a life that was in keeping with our way, with our father's profession, and our little girl should live that life with us.	0	0	0	0	0	1	0	0	0
1958	vol008	nr001	THE ISLAND (Diabetes was not an obstacle	My husband and I knew that we were in God's hands wherever we went and that there were great doctors and good hospitals everywhere! And insulin was available wherever we went. All we needed was faith in God and a little intelligence and certainly a few good people willing to help us.	0	0	0	0	0	1	0	0	0
1958	vol008	nr001	THE ISLAND (Diabetes was not an obstacle	But I was still scared. Especially because I had never seen the inside of an airplane, let alone flown in one. And now I was about to fly over 11,000 km. That's no small feat for a first trip in an airplane!	0	0	0	0	0	1	0	0	0

Figure 1: An example from the annotated subset of "Diabetiker Journal"

each genre, we aggregate the available data from all datasets.

Each model is fine-tuned on all the configurations resulting in 48 fine-tuned models. A random 10% portion of each training set with stratification by label is used for validation. For an intra-dataset evaluation of the models, we use a test set described in Table 3. This test set is shared by all configurations within the same dataset type (CORE, FTD, UDM, merged). Pre-processing of the data includes removal of web addresses, emails, XML-tags, and emoji.

4.2 ActDisease Dataset

The ActDisease Dataset is a private dataset currently undergoing digitization (Aangenendt et al., 2024). At this stage, the digitization process is completed for two languages, Swedish and German, and partially completed for French. In this paper, we focus on the magazines issued in German and Swedish. German data covers the 1875–1990 period and Swedish data—1938–1990. The dataset contains 64863 issues with a total size of 112M tokens.

Preprocessing for annotation. The dataset is initially represented as the XML output of the Optical Character Recognition (OCR) engine ABBYY Finereader 14[4] for each page. To facilitate the annotation process, we use the following procedure to extract continuous text fragments under each title in each issue.

For each recognized paragraph, font attribute patterns of its lines (size, font type, font size, bold, italic) are collected from the OCR output. Consequent lines with the font attribute pattern are merged into paragraphs and paragraphs containing only non-words are dropped. Content pages are identified using regex and the titles are collected.

Each issue's paragraphs are represented as a sequence of font sizes attributes and are clustered using GaussianHMM, which is a well-known method in speech pattern recognition (Bilmes, 2008). Clusters corresponding to titles are identified using fuzzy string matching: MinHash locality sensitive hashing (Broder, 1997) and TheFuzz[5], a tool based on the Levenstein distance. The cluster that follows the title cluster and contains the longest sequence of uniform font attribute patterns is consider to be part of text under this title. It is added to the dataset with year, volume, issue number and title as descriptors. The clusters that precede or follow this text cluster are added with the same title if they contain the same font attribute pattern.

Annotation. Annotation files (Numbers spreadsheets) are produced for two periodicals: the Swedish "Diabetes" and the German "Diabetiker Journal". To increase variation in content, we select the first and mid-year issues in each year. An example of annotations for several translated paragraphs from the "Diabetiker Journal" are shown in Figure 1. The annotation was performed by 4 historians and 2 computational linguists either native or proficient in Swedish and German. At least 2 annotations were collected for each paragraph. The average kappa agreement is 0.7. The final dataset includes only those paragraphs, for which at least two annotators agree.

Sampling Strategy. The annotated dataset is divided into a training set (1182 paragraphs) and a held-out set (552 paragraphs), with stratification based on labels. The distribution of paragraphs across languages and genres in these sets is illustrated in Figures 2 and 3.

For few-shot experiments, models are trained on six different training set sizes: 100, 200, 300, 400, 500, and 1182 instances. The subsets are randomly sampled from the training set, ensuring each is balanced by label. The held-out set is further divided into a validation set and a test set, each containing an equal number of instances and preserving label balance. The legal and news categories are

[4] https://www.abbyy.com/company/news/abbyy-finereader-14-pdf-solution/

[5] https://github.com/seatgeek/thefuzz

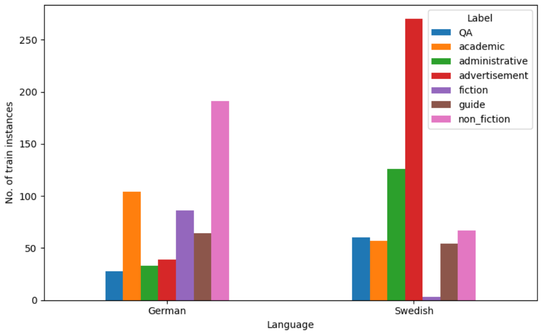


Figure 2: Genre distribution in languages in the training sample of the ActDisease Dataset

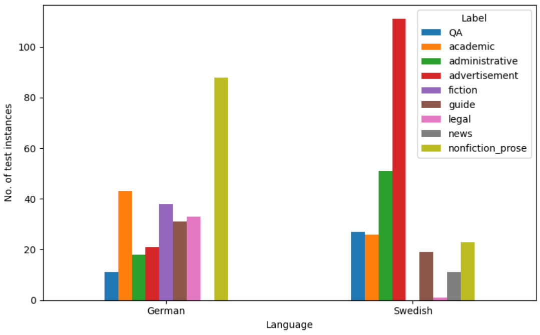


Figure 3: Genre distribution in languages in the held-out sample of the ActDisease Dataset

excluded from these experiments due to insufficient training data.

For zero-shot experiments, the entire test set is used as the test set.

5 Experimental Setup

The main goal of this study is to identify the optimal training dataset configurations and fine-tuning strategies for aligning with the annotated genre labels. We explore two scenarios: zero-shot prediction, where models predict genres using existing web genre datasets without seeing the target data, and few-shot vs. semi-supervised few-shot training. In the semi-supervised scenario, we examine if pre-training the models for MLM on the entire ActDisease Dataset improves their few-shot prediction performance.

5.1 Models

For fine-tuning, we utilize pre-trained base versions of mBERT, XLM-RoBERTa and historical multilingual model hmBERT on the AGI datasets. BERT-like models have been extensively used in the previous work for web register and genre classification (Lepekhin and Sharoff, 2022; Kuzman and Ljubešić, 2023; Laippala et al., 2023). XLM-RoBERTa outperformed mBERT on the XNLI benchmark (Conneau et al., 2020) and has recently been successfully applied for web genre classification in (Kuzman et al., 2023).

hmBERT is relevant for this work, since it is pre-trained on a large corpus of historical newspapers. The Swedish portion spans publications from 1900 to 1910, while the German dataset provides good coverage of the 19th and 20th centuries.

mBERT is used for comparison with hmBERT since both are based on BERT, while XLM-RoBERTa is not directly comparable.

5.2 Zero-Shot Prediction

In historical NLP, in-domain training data is often unavailable. To address this, we fine-tune our models on each out-of-domain AGI training dataset configuration individually, as well as on a merged version that combines all datasets. We begin by evaluating the classifiers' predictions on their respective native test sets. Since we map the original labels to our genre categories, this change in genre representation is likely to affect the models' inference.

Following this, we perform zero-shot prediction on the ActDisease Dataset's test set and compare the results to a baseline. This scenario is cross-lingual for the FTD and X-GENRE datasets because they lack German and Swedish instances. For the UDM and CORE datasets, the scenario is partially cross-lingual: UDM includes Swedish instances in guide, fiction, and administrative categories, and German - in news; CORE contains a small number of Swedish instances in the guide category.

Baseline. We use the state-of-the-art classifier of web genres - X-GENRE (Kuzman et al., 2023) as a baseline. We consider the predictions on the most similar labels that can be directly mapped to ours: Instruction (mapped to: guide), Legal, News, Promotion (advertisement), Prose/Lyrical (fiction).

5.3 Few-shot and semi-supervised training

In this experiment, we explore a scenario with a limited number of annotated training examples. We train the models on datasets of different sizes, ranging from 100 examples up to the full training dataset of 1182 paragraphs. The training is conducted in two modes: with and without an initial phase of MLM pre-training on the entire ActDisease Dataset.

	FTD		CORE		UDM		merged	
	ACC	Macro-F1	ACC	Macro-F1	ACC	Macro-F1	ACC	Macro-F1
hmBERT	0.66	0.68	0.75	0.75	0.68	0.69	0.67	0.68
mBERT	0.88	0.88	0.76	0.77	**0.82**	**0.83**	0.80	0.79
XLM-RoBERTa	**0.91**	**0.91**	**0.78**	**0.78**	**0.82**	**0.83**	**0.83**	**0.83**

Table 4: Intra-dataset evaluation of the classifiers. Average scores over the models trained on different dataset configurations.

		QA	academic	administrative	advertisement	fiction	guide	legal	news	nf_prose
	X-GENRE	-	-	-	0.69	0.39	0.59	0.66	0.08	-
FTD	hmBERT	-	0.37	-	0.56	0.43	0.33	**0.9**	0.38	0.47
	mBERT	-	0.61	-	0.62	0.40	0.47	**0.82**	0.34	0.54
	XLM-RoBERTa	-	0.57	-	0.74	0.49	0.57	**0.89**	0.28	0.56
CORE	hmBERT	0.1	0.45	-	0.07	0.41	0.23	**0.80**	0	0.20
	mBERT	0.18	0.48	-	0.10	0.32	0.26	**0.80**	0	0.34
	XLM-RoBERTa	0.35	0.50	-	0.11	0.46	0.30	**0.84**	0.07	0.33
UDM	hmBERT	0.1	0.04	**0.43**	-	0.27	0.26	0.09	0.01	0.03
	mBERT	0.16	0.25	0.25	-	0.17	0.29	0.16	0.04	0.01
	XLM-RoBERTa	**0.53**	0.21	0.30	-	0.30	0.31	0.14	0.05	0.08
merged	hmBERT	0.10	0.18	0.24	0.22	0.36	0.18	0.19	0.01	0.17
	mBERT	0.05	0.15	0.18	0.23	0.19	0.17	0.21	0.02	0.09
	XLM-RoBERTa	0.43	0.27	0.14	0.34	0.40	0.28	0.45	0.04	0.14

Table 5: Zero-shot per-category F1 scores averaged across dataset configurations. The highlighted values indicate cases where the highest average F1 performance for a certain category does not result from systematic overprediction of this category by the classifiers, as verified through our analysis of confusion matrices.

6 Results

6.1 Intra-Dataset Evaluation of the Classifiers

The evaluation of models fine-tuned on web genre datasets is presented in Table 4, where they are assessed against their corresponding native test sets. The results are averaged across dataset configurations. XLM-RoBERTa shows the best performance across all datasets. For UDM, both XLM-RoBERTa and mBERT greatly outperform hmBERT. hmBERT achieves the lowest scores on all datasets, which is expected in view of the nature of its historical training data. The best genre prediction capacity is observed on the FTD dataset with XLM-RoBERTa.

The scores of mBERT and XLM-RoBERTa on the CORE dataset with our genre mapping are noticeably lower than on other datasets. Moreover, during fine-tuning, overfitting occurs earlier for CORE (from the 3rd epoch on average) than for UDM or FTD (from the 5th epoch on average). This performance may indicate that our genre mapping for this dataset is inappropriate.

The results for the merged dataset are not surprising in view of the performance of the CORE dataset, since CORE instances dominate in the merged training data. Similarly to CORE, overfitting occurs earlier for the models trained on the merged dataset (from the 3rd epoch on average).

6.2 Zero-Shot Inference

Table 5 presents the zero-shot inference results (F1 scores) for various genres, averaged across different dataset configurations. Since each AGI dataset contains only a subset of the genres, it is not possible to directly compare the overall performance metrics of the classifiers. Instead, we evaluate the performance for each genre separately and analyze the confusion matrices[6] to mitigate potential biases.

In general, our analysis indicates that models trained on the FTD dataset configurations perform better with our genre mapping compared to models trained on other datasets. It also suggests that merging datasets necessitates a different approach to achieve optimal results.

Upon close examination of the results from models trained on the UDM dataset, we observe a class-specific bias. Despite applying downsampling, models trained on all dataset configurations tend to overpredict the news category. The average

[6] Appendix C contains confusion matrices that showcase the trends observed in zero-shot inference

accuracy of these models remains below 0.5. The proportion of news instances in the dataset configurations is consistent with other downsampled categories like academic, fiction, and legal, averaging around 15%. However, the news category includes the highest number of Germanic instances, most of which are in German, potentially explaining this bias.

Interestingly, the administrative category, when classified by hmBERT, is less affected by this bias compared to other categories. Furthermore, hmBERT correctly predicts on average 25 out of 69 administrative instances, outperforming mBERT (11) and XLM-RoBERTa (14). A possible explanation for this could involve two factors: 1) hmBERT's pre-training on historical newspapers, which extensively used the report genre—characterized by near-verbatim chronological documentation of meetings and events (Bødker, 2020), and 2) textual similarity between patient organization meeting records and European parliamentary meeting minutes in the UDM training set.

On a different note, XLM-RoBERTa shows superior performance in the QA category, averaging 16 correct predictions out of 38 instances, while mBERT and hmBERT achieve only 4 and 2 correct predictions, respectively.

Classifiers trained on FTD and CORE show strong performance in predicting the legal category with no biases detected in our confusion matrix analysis with respect to this category.

Table 6 illustrates the average impact of different configuration options on fine-tuning for each dataset. For the FTD dataset, additional balancing based on the original labels or filtering by language family does not enhance performance. Although the CORE dataset is predominantly English, the inclusion of a small number of Finnish and French instances slightly diminishes its performance. For the UDM dataset, the presence of other language families and balancing generally improve performance in terms of macro F1.

In summary, for cross-lingual zero-shot prediction, training on the FTD dataset using our genre mapping is more effective than training on CORE or UDM, or using a pre-trained multilingual genre classifier.

6.3 Few-Shot Inference

Fig 4 shows the trends in performance of the models on the multilingual historical test set after being fine-tuned on training sets of various sizes. Further pre-training with a MLM objective is clearly advantageous. F1 keeps increasing with the number of training instances but is still below 0.8 with 1182 training instances for all models. hmBERT-MLM outperforms XLM-RoBERTa-MLM and mBERT-MLM by a small margin. mBERT-MLM is very close to hmBERT-MLM with dataset size 500. For the dataset size of 400, a decline in performance is observed for XLM-RoBERTa, XLM-RoBERTa-MLM, and hmBERT-MLM. Particularly for XLM-RoBERTa, fine-tuning on this dataset portion greatly increases confusion between QA and academic, advertisement and guide, as well as nf_prose and fiction. For the models with MLM-pretraining, the confusion is less pronounced. Further investigation is needed to understand the un-

	configuration	*F1*
FTD	[B2]	0.51
	[B1]	**0.56**
	[G+]	0.52
	[G-]	**0.56**
CORE	[B2]	0.32
	[B1]	0.32
	[G+]	**0.32**
	[G-]	0.31
UDM	[B2]	**0.19**
	[B1]	0.18
	[G+]	0.14
	[G-]	**0.22**

Table 6: Macro F1 scores of models trained on the AGI datasets averaged by configuration settings. [B2] means balanced by two levels: our label and original dataset labels. [B1] means balancing by our labels only. [G+] means the filtering by language family is performed and only Germanic languages are present in the dataset. [G-] is for the case when all language families are included in the dataset

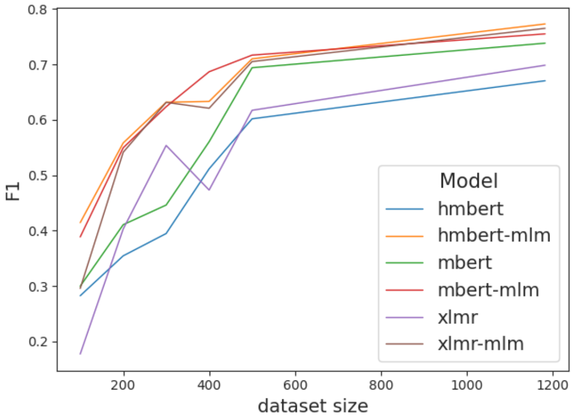


Figure 4: Performance of the models in a few-shot setting with and without MLM fine-tuning.

MODEL	XLMR		XLMR-MLM		hmBERT		hmBERT-MLM		mBERT		mBERT-MLM	
SIZE	500	1182	500	1182	500	1182	500	1182	500	1182	500	1182
QA	0.61	0.76	**0.77**	**0.84**	0.55	0.73	0.71	0.76	0.62	0.72	0.73	0.78
academic	0.63	**0.81**	0.75	**0.81**	0.62	0.76	0.70	0.78	0.75	0.77	**0.77**	**0.81**
administrative	0.78	0.84	0.79	0.84	0.70	0.82	0.77	**0.86**	**0.82**	0.82	0.80	**0.86**
advertisement	0.81	**0.93**	**0.90**	**0.93**	0.84	0.91	0.89	**0.93**	0.87	0.92	0.87	0.92
fiction	0.49	0.05	0.55	0.34	0.53	0.10	0.60	**0.51**	0.58	0.47	**0.62**	0.33
guide	0.67	0.76	**0.73**	**0.79**	0.52	0.58	0.64	0.72	0.68	0.68	0.67	**0.79**
nf_prose	0.30	0.72	0.42	0.77	0.42	0.74	**0.62**	**0.80**	0.51	0.75	0.53	0.78
accuracy	0.63	0.78	0.72	0.81	0.63	0.76	**0.73**	**0.82**	0.71	0.78	**0.73**	0.81
macro_F1	0.61	0.69	0.70	0.76	0.60	0.67	**0.71**	**0.77**	0.69	0.73	**0.71**	0.75

Table 7: Per-category F1 and overall metrics achieved by pre-trained models in a few-shot setting (with and without MLM fine-tuning) for two dataset sizes: 500 and 1182 training instances.

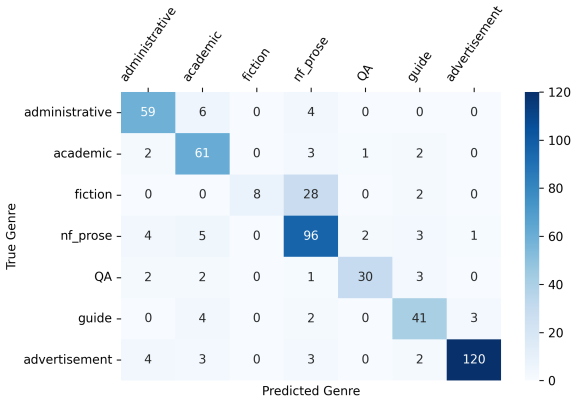

	administrative	academic	fiction	nf_prose	QA	guide	advertisement
administrative	59	6	0	4	0	0	0
academic	2	61	0	3	1	2	0
fiction	0	0	8	28	0	2	0
nf_prose	4	5	0	96	2	3	1
QA	2	2	0	1	30	3	0
guide	0	4	0	2	0	41	3
advertisement	4	3	0	3	0	2	120

Figure 5: XLM-Roberta-MLM classification results with full-sized training dataset

derlying causes.

Table 7 provides further details on the label-wise F1 scores for dataset sizes 500 and 1182 (light grey columns). Although hmBERT outperforms other models in terms of overall accuracy and F1 score, label-wise F1 scores show that this is largely due to stronger prediction of fiction and nonfiction prose and a less drastic drop in fiction with the dataset size increase.

An analysis of confusion matrices, such as the one depicted in Figure 5 for XLM-Roberta-MLM, reveals that nonfictional prose is frequently overpredicted when the model is trained on the entire training dataset. This overprediction indicates that the fiction and nonfictional prose categories may be becoming increasingly similar, causing greater confusion for the classifiers and resulting in higher misclassification rates.

In contrast, hmBERT-MLM exhibits a lower susceptibility to this confusion compared to other models, suggesting it is better at distinguishing between these categories even as they become more similar.

Genre identification in this context is particularly challenging because all genres are confined to a specific domain: patient organizations' magazines focused on diabetes. This means that both fictional and (auto)biographical narratives frequently revolve around the experiences of diabetes patients, and are likely to share themes and narrative structures.

Among the models not further pre-trained on the ActDisease Dataset, mBERT achieves a surprisingly strong macro F1 score compared to the others.

Additional pre-training with a MLM objective enhances the quality of the few-shot learning for the three considered models. It results in considerable gains: on average 18.5% across models trained on all dataset sizes. The greatest average increase in macro F1 is observed for hmBERT-MLM (24% as opposed to 14.5% for mBERT-MLM and 16.9% for XLM-RoBERTa). For XLM-RoBERTa and hmBERT, the effect of over-fitting in the fiction category with the full-size dataset becomes less pronounced in the MLM-fine-tuned models. It is not the case in the mBERT though.

7 Conclusions

In this work, we address an underexplored problem of genre classification for historical magazines. First, we show that with a custom genre scheme based on the dataset properties it is possible to successfully leverage the categories available in the modern datasets in cross-domain and cross-lingual zero-shot prediction. Our analysis reveals that models trained on the FTD dataset configurations achieve better alignment with our genre mapping compared to those trained on other datasets.

Next, we highlight the advantages of few-shot learning using a small set of annotated instances. Even a limited annotated sample from the same data source greatly enhances genre classification performance on our historical test dataset. Furthermore, we find that prior MLM fine-tuning substantially improves few-shot learning across all models, with particularly strong gains for historical multilingual BERT.

For future work, we aim to expand annotation efforts to include new genre categories and languages (English and French). Once sufficient annotations are available, we will also explore monolingual few-shot experiments to compare the performance of monolingual and multilingual large language models on this task. In addition, we plan to investigate how linguistic similarities between training and test genre data are related to the classification performance.

8 Limitations

Our study acknowledges several limitations that should be addressed in future research. While we are actively working on expanding the dataset, the size of our annotated dataset in these experiments is relatively small, which may restrict the generalizability and robustness of our findings. A larger corpus would provide more comprehensive training data and potentially lead to more reliable model performance.

Additionally, the annotated dataset exhibits some degree of imbalance across different genres. This imbalance could introduce biases during the training process, affecting the overall performance and fairness of our models.

Moreover, due to the scarcity of annotated instances for individual languages, we do not cover the monolingual few-shot setup. It would involve fine-tuning language-specific pre-trained models for MLM and then performing few-shot training. However, in this scenario, much less data would be used for MLM-fine-tuning.

Lastly, our pre-processing for annotation extracts mostly clean paragraphs with low OCR error rate. In real-life cases, the paragraphs are often noisy and suffer from poor OCR, especially in case of unusual fonts or layouts. We plan to address this in the future research.

Acknowledgments

This research is funded by the European Union (ERC, ActDisease, ERC-2021-STG 10104099). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

We would like to thank the **Centre for Digital Humanities and Social Sciences** at Uppsala University for providing us with the computational resources for training and evaluating our models. We are especially grateful to Dr. Maria Skeppstedt, Dr. Julia Reed, Dr. Andrew Burchell, and Gijs Aangenendt for their invaluable contributions to annotation, category definition, and mapping, as well as their feedback on the paper. We sincerely thank Dr. Maria Skeppstedt and the anonymous reviewers for their insightful comments and suggestions, which greatly enhanced the quality of the paper.

References

Gijs Aangenendt, Maria Skeppstedt, and Ylva Söderfeldt. 2024. Curating a historical source corpus of 20th century patient organization periodicals. In *Proceedings of the Huminfra Conference (HiC 2024)*, pages 76–82.

Jeff Bilmes. 2008. *Gaussian Models in Automatic Speech Recognition*, pages 521–555. Springer New York, New York, NY.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of SEQUENCES 1997*, pages 21–29. IEEE. Cat. No.97TB100171.

Marcel Broersma. 2010. Journalism as a performative discourse: The importance of form and style in journalism. In Verica Rupar, editor, *Journalism and Meaning-Making: Reading the Newspaper*, pages 15–35. Hampton Press, Cresskill.

Marcel Broersma and Frank Harbers. 2018. Exploring machine learning to study the long-term transformation of news. *Digital Journalism*, 6(9):1150–1164.

Henrik Bødker, editor. 2020. *Journalism History and Digital Archives*, 1st edition. Routledge.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nan Z. Da. 2019. The computational case against computational literary studies. *Critical Inquiry*, 45(3):601–639.

Vera Danilova and Sara Stymne. 2023. UD-MULTIGENRE – a UD-based dataset enriched with instance-level genre annotations. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 253–267, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Marta Hanson. 2022. Epistemic genres as a conceptual tool in the history of chinese medicine. *Chinese Medicine and Culture*, 5(1):1–8.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain. Association for Computational Linguistics.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: a survey. *Language Resources and Evaluation*.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of english. *Language Resources and Evaluation*, 57(3):1045–1079.

Mikhail Lepekhin and Serge Sharoff. 2022. Estimating confidence of predictions of individual classifiers and TheirEnsembles for the genre classification task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5974–5982, Marseille, France. European Language Resources Association.

F. Moretti. 2000. Conjectures on world literature. *New Left Review*, 1(1):54–68.

Gianna Pomata. 2014. The medical case narrative: distant reading of an epistemic genre. *Literature and medicine*, 32(1):1–23.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 183–191, Online. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. *CoRR*, abs/2205.15575.

Serge Sharoff. 2021. Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3(1):1–32.

A Genre Annotation Guidelines

A.1 Genre Definitions

Academic reports about academic research or explains complex scientific ideas or discoveries in an accessible way.

Author : medical professionals, researchers

Target audience : physicians, researchers, patients, or other readers of the periodical.

Features : 1) high density of specialist language including domain-specific terms (e.g., "coronary angiography") and research terms (e.g., "experiment", "approach", "results", "method"); 2) references to academic works; 3) factual, often impersonal, narrative about an observation of a process (experiment/treatment/chemical reactions) and its outcomes.

Subgenres : academic article, academic report, popular science article

Administrative reports about the activities or discusses plans of the patient organization.

Author : directive authorities or members

Target audience : members, directive authorities of another organization/authority, politicians

Features : 1) presence of named entities referring to the organization and its directive members; 2) terms such as "annual meeting", "financial report", "association report", "association activities", "association meeting"; 3) detailed chronological reporting;

Subgenres : meeting minutes, financial reports, annual reports, editorial information, official correspondence and petitions, announcements

Advertisement promotion of products and services with intent to sell them, e.g.: sweeteners, injectors, alcohol, yoga, courses for nurses, lotteries. These texts aim to create awareness of brands, products, services, and ideas, as well as to persuade the public to respond in a certain way toward what is advertised.

Subgenres : advertisement, promotion, invitation

Guide recommends, provides advice or instructions for step-by-step implementation to achieve a certain goal or solve a problem related to health, legal issues or other. It can be one step-action or more.

Author : directive authorities, members of the organization, medical doctors, dieticians, patients, consultants

Target audience : members or other readers of the periodical

Features : 1) imperative modality expressed with auxiliary verbs such as "should", "must"; 2) itemized lists of actions; 3) addresses the reader in 2nd person plural; 4) chronological order; 5) presence of expressions similar to "It is recommended to"/"We recommend you to" or "It is advisable"/"We advise you"

Subgenres : dietary advice, physical exercise instructions, recipe, procedural instructions, application guidelines

Fiction aims to entertain the reader, gives reading pleasure, engages the reader emotionally.

Author : fiction authors

Target audience : members or other readers of the periodical

Features : 1) presence of imaginary elements, such as invented characters, events, worlds; 2) dense use of creative language such as tropes; 3) emotional engagement; 4) can include dialogue of characters

Subgenres : poems, short stories, humor, myths, novel, novella

Legal explains or informs about law.

Subgenres : contracts, terms and conditions

News report about recent events. Contains short factual text announcing an event with no analysis or literary narrative, not a long-read.

Subgenres : daily news reports

Nonfictional prose narrates/reports about events/experiences from personal life or represents a neutral description of cultural phenomena or history.

Author : members of the organization, patients

Target audience : members or other readers of the periodical

Features : 1) first-/third-person narrative; 2) chronological perspective; 3) references to time; 4) factual or opinion; 5) language is not rich in tropes; 6) informal or neutral language;

Subgenres : auto(biography), memoire, travel note, personal letter, opinion essay, cultural article, documentary prose

QA is text structured in a question-answer format, for example, questions from members and answers from medical professionals. Most frequently corresponds to the questions and answers section of the magazine.

A.2 Criteria for Genre Assignment

We base the categorization on concepts shared by these sources that closely align with the idea of communicative purpose. Although communicative purpose is itself a complex and multilayer concept, it has often been considered a key characteristic feature for genre identification and categorization.

We perform classification on the paragraph level. Each paragraph is part of a column text under a certain title. Title often indicates what type of text all the underlying paragraphs belong to. E.g., the “Våra lokalföreningar” will indicate that the following text discusses organizational activities.

The annotator is given a table where each row includes a paragraph with its identifiers (journal, year, volume, issue number), the corresponding title and empty genre category columns. The annotator should place his hard assignment (1) in the corresponding column in front of each paragraph.

B Fine-tuning Settings

The following hyper-parameters were used in our experiments for fine-tuning in the zero shot and few-shot settings:

- Number of epochs: 10
- Learning rate: $1e-5$
- Batch size: 8
- Weight decay: 0
- Maximum sequence length: 512

Other settings are set to default values in the Huggingface’s Trainer[7].

For the MLM fine-tuning, we used the official script `run_mlm.py` available in the transformers GitHub[8] with batch size equal to 8.

[7] https://huggingface.co/docs/transformers/main_classes/trainer

[8] https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling

C Classifier Performance in Zero-Shot Evaluation

The following figures illustrate zero-shot performance of classifiers fine-tuned on existing datasets for web genre and register classification when applied to our historical test dataset, as discussed in Section 6.2.

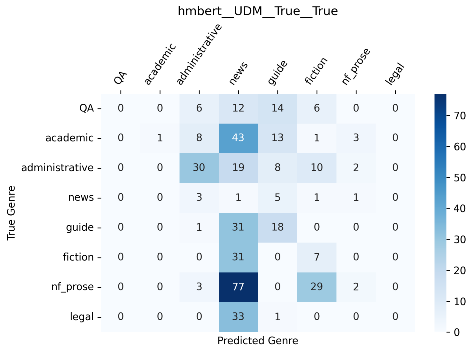


Figure 6: Classification results for hmBERT fine-tuned on UDM (B2 G+). The classifier recognizes 43% of paragraphs in the administrative genre.

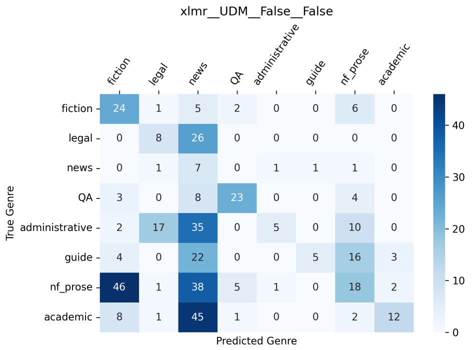


Figure 7: Classification results for XLM-RoBERTA fine-tuned on UDM (B1 G-). The classifier recognizes 60% of paragraphs in the QA genre.

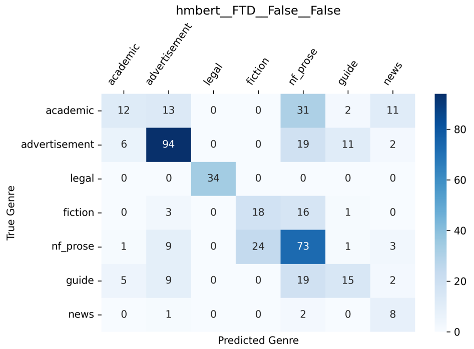


Figure 8: Classification results for hmBERT fine-tuned on FTD (B1 G-). The classifier recognizes 100% of paragraphs in the legal genre.

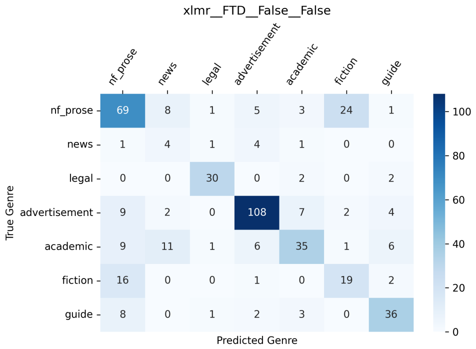


Figure 9: Classification results for XLMR-RoBERTA on FTD (B1 G-). The classifier recognizes 88% of paragraphs in the legal genre.

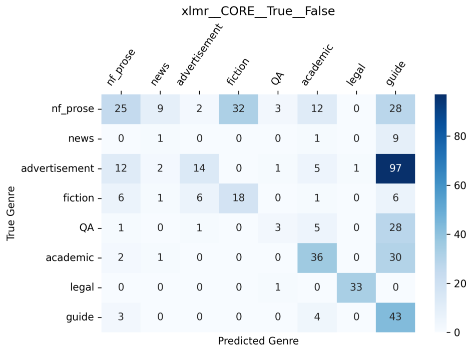


Figure 10: Classification results for XLM-Roberta fine-tuned on CORE (B1 G+). The classifier recognizes 97% of paragraphs in the legal genre.

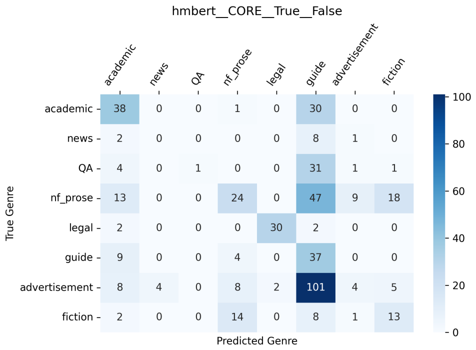


Figure 11: Classification results for hmBERT fine-tuned on CORE (B1 G+). The classifier recognizes 88% of paragraphs in the legal genre.