Detecting Sequential Genre Change in Eighteenth-Century Texts

Zhang, Jinbin

CEUR-WS.org
2022-12-12

Zhang , J , Ryan , Y C , Rastas , I , Ginter , F , Tolonen , M & Babbar , R 2022 , Detecting Sequential Genre Change in Eighteenth-Century Texts . in F Karsdorp , A Lassche & K Nielbo (eds) , Proceedings of the Computational Humanities Research Conference 2022 . CEUR Workshop Proceedings , vol. 3290 , CEUR-WS.org , Aachen , pp. 243-255 , Computational Humanities Research Conference , Antwerp , Belgium , 12/12/2022 .

http://hdl.handle.net/10138/351519

Detecting Sequential Genre Change in Eighteenth-Century Texts

Jinbin Zhang[1], Yann Ciarán Ryan[3], Iiro Rastas[4], Filip Ginter[4], Mikko Tolonen[3] and Rohit Babbar[1]

[1]*Aalto University, Finland*
[3]*University of Helsinki, Finland*
[4]*TurkuNLP, University of Turku, Finland*

Abstract

Machine classification of historical books into genres is a common task for NLP-based classifiers and has a number of applications, from literary analysis to information retrieval. However it is not a straightforward task, as genre labels can be ambiguous and subject to temporal change, and moreoever many books consist of mixed or miscellaneous genres. In this paper we describe a work-in-progress method by which genre predictions can be used to determine longer sequences of genre change within books, which we test out with visualisations of some hand-picked texts. We apply state-of-the-art methods to the task, including a BERT-based transformer and character-level Perceiver model, both pre-trained on a large collection of eighteenth century works (ECCO), using a new set of hand-annotated documents created to reflect historical divisions. Results show that both models perform significantly better than a linear baseline, particularly when ECCO-BERT is combined with tfidf features, though for this task the character-level model provides no obvious advantage. Initial evaluation of the genre sequence method shows it may in the future be useful in determining and dividing the multiple genres of miscellaneous and hybrid historical texts.

Keywords

BERT, text classification, genre change, ECCO, Perceiver

1. Introduction

Thinking about large-scale development of early modern public discourse through the use of structured data is an exciting opportunity as was established by Moretti some time ago. [19] Besides the use of already available bibliographic data for "distant reading", a useful further element is to use unstructured textual databases as source material for the creation of new structured data on fields that are currently poorly available. [15] One such classification field is genre. Readily available genre information is often sporadic, but the opportunities to use it – especially when we think that many documents are composed of several sequential genres – can open a new window to the development of public discourse. With better structured data,

CHR 2022: Computational Humanities Research Conference, December 12–14, 2022, Antwerp, Belgium
jinbin.zhang@aalto.fi (J. Zhang); yann.ryan@helsinki.fi (Y. C. Ryan); iitara@utu.fi (I. Rastas); figint@utu.fi (F. Ginter); mikko.tolonen@helsinki.fi (M. Tolonen); rohit.babbar@aalto.fi (R. Babbar)
yann-ryan.github.io (Y. C. Ryan)
0000-0001-8186-8677 (J. Zhang); 0000-0003-1878-4838 (Y. C. Ryan); 0000-0003-2892-8911 (M. Tolonen); 0000-0002-3787-8971 (R. Babbar)

we will be able to study the systematization of particular genres in a new manner and take a fresh look on authorship and the relevance of publisher networks.

Much work in literary history and the history of the book has relied on the analysis of generic categories (for examples see [20, 33, 34, 35, 2, 19]). Computational genre classification is a complex problem. Two key reasons are that genre divisions change over time, and not every book can be unambiguously assigned a single genre label. Existing methods for genre detection often assume each text or pre-defined chunk such as a chapter or section can be classified as a single genre or a distribution of genre probabilities [7, 38, 6], which does not reflect the reality of many eighteenth century texts. One important exception to this is the page-level classification of Underwood et al. [36], subsequently used to detect sequences of genre using a hidden Markov model. [37]

This paper describes a number of improvements to existing methods: first, rather than relying on existing modern or broad classification systems, we use a newly-created training set of documents, with a custom-designed, domain-specific taxonomy which attempts to balance pragmatism with capturing meaningful and fine-grained eighteenth-century organisational categories. Second, we use a BERT transformer model which has been specifically trained on eighteenth century texts, which performs significantly better than base BERT, and third, we propose a method by which we hope this fine-grained classification can be used to represent books as sequences and combinations of genres.

We report on and compare results from a number of classifiers: a document-level classifier that uses only one BERT input segment for each document (ECCO-BERT-Seq), a classifier for text chunks, which can also be aggregated on a document-level (ECCO-BERT-Chunk), and a character-level Perceiver model using the same input as ECCO-BERT-Seq.[1] The BERT model [11] has achieved great improvements on various modern language datasets in comparison to previous deep learning methods. Recently, there have also been some models which are pre-trained on historical corpora of different languages [21, 16, 39], and pre-trained language models are also used in the historical domain, such as predicting the year [21], named entity recognition [16, 13, 1, 27] and emotion analysis. [26, 25] We also face some challenges from OCR recognition errors [10, 29] when using pre-trained models for historical data.

2. The ECCO Dataset

The data used both for model training and for predictions comes from Eighteenth Century Collections Online (ECCO). ECCO is a set of 180,000 digitised documents published originally in the eighteenth century, created by the software and education company Gale. [5] These digitised images have been converted into readable text data using Optical Character Recognition (OCR). Despite its size, a recent study comparing ECCO to the English Short Title Catalogue (ESTC) has highlighted significant gaps and imbalances[32], and the ESTC itself is known to be incomplete. [22] These attributes, and the impact of them on several downstream tasks,

[1] In this paper the words 'book' and 'document' have distinct meanings. 'Book' is used to denote an edition of a physical book, for example 'there are over 400,000 books listed in the English Short Title Catalogue'. 'Document' by contrast, is reserved for a single text document as used for data for the classification method and other tasks. Not all documents in the ECCO data map to a single book, and vice-versa.

have been covered in detail in previous papers [30, 8, 14] and are just briefly outlined here. First, the distribution of documents in ECCO is uneven and skewed towards the end of the century and second, the OCR contains significant noise and errors. Additionally, not all texts are in the English language, and many are reprints of works published in earlier centuries. The former have been excluded but the latter are retained for our training and test data. Despite these caveats, ECCO is the largest and most complete source we have for eighteenth-century text data. Though it has its own institutional history and biases, it is complete enough that it contains not only the more 'important' or 'literary' genres, nor is it focused solely on canonical works. Its data and digitised images are used extensively, forming the basis of many scholarly enquiries and research questions. [31]

3. Data Annotation

Key to the work leading up to this paper was to create a usable training set of documents annotated with genre labels. We began with a sample set of book records and a set of preliminary genre labels. These books were then labelled by two annotators with domain expertise. At this stage, we revisited the labels, and made some adjustments to those which had particularly low inter-annotator agreement. Once the set of genre labels had been finalised, we annotated a large set (5,672 individual works, which correspond to 37,574 known editions, of which 30,119 correspond to ECCO documents) with genre information. After this second round, we again checked for inter-annotation agreement, coming to a consensus following a discussion of each disagreement. The eventual 43 fine-grained categories were then collapsed into main categories for some of the classification tasks. These book labels were then mapped to the equivalent ECCO document IDs. The final set of labels are given in appendix A.

Existing categorical distinctions were either too broad (for example fiction and non-fiction) or too fine-grained (for example the many historical literary divisions, particularly poetic) for our needs. Our categories attempt to reflect the divisions as found in contemporary sources such as catalogues. [17] Additionally, they are closely related to the divisions used by modern domain experts writing on the history of the book, for example the chapters of the highly-regarded edited collection *Books and their Readers in Eighteenth-Century England*, which contains chapters organised along similar divisions to our own. [23, 24] We note that other recent attempts to categorise eighteenth century book genres use a similar system of division. [18] The selection is intended to provide useful genre categorisation for scholarly inquiry into book history and book production. The selection was also pragmatic, with the aim of ending up with a manageable number of genres, for example so that each class had enough data for the training and test sets. They were also made with particular questions in mind, which we hoped would help us to analyse works of Scottish Enlightenment thought, for instance helping to distinguish patterns within scientific or philosophical publishing.

4. Method

In this section, we introduce the pre-trained ECCO-BERT model, fine-tuning models and baselines.[2] We denote the training dataset as $\{(X_i, Y_i)\}_{i=1}^{N}$, where X_i is the book, and Y_i is the genre of X_i. Our goal is to learn a function $f(X_i)$ to predict the genre for book X_i or the genre of a chunk in book X_i.

4.1. Multi-granular Classification with ECCO-BERT

ECCO-BERT [21] is a pre-trained language model trained on the ECCO dataset, the configuration of which is the same as the bert-base-cased model [11] except for the vocabulary size. The model is pre-trained with a masked language modelling task, as well as a next sentence prediction task. The fine-tuned ECCO-BERT consists of two parts, one is the transformer encoder and the other is the linear layer on the top of mean pooling output of the encoder, which scores different genres. The Transformer model architecture on which the model is based can accept inputs up to a relatively short maximum length, in the ECCO-BERT case the standard maximum of 512 input tokens applies. Inputs longer than this maximum length need to be split into chunks.

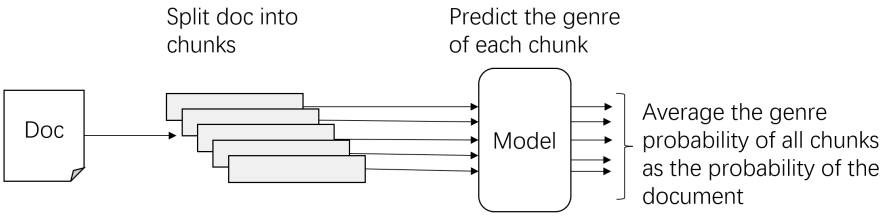

Figure 1: The inference process of ECCO-BERT-Chunk. The document is torn into chunks without overlap. The model scores each chunk and averages the probability of each chunk as the final prediction.

Because we want the training and prediction of the model to take into account the full information of the document, a document is torn into different chunks of 510 tokens each to train the model and predict results, since the maximum input size of ECCO-BERT is 512 tokens (510 input tokens and 2 special tokens expected by the model). For training the model, we assume that each chunk has the same genre as the document, and the model is trained with the resulting (chunk, label) pairs. During the inference procedure, we first split the document into chunks. The fine-tuned model then scores each chunk; the predicted genre probability of the document is the average of all chunks' probability. The inference process is shown in Figure 1. We call this model ECCO-BERT-Chunk. For comparison, we also train a model conditioned only on the first 510 sub-words of the document as input, which is denoted as ECCO-BERT-Seq.

Although the ECCO-BERT-Chunk model considers all chunks to make the final judgment, its prediction process is very slow since a book often contains a lot of chunks. At the same time, the much faster ECCO-BERT-Seq is only conditioned on the first 510 sub-words, so it might lose some important information of other parts in the book. To solve this problem, we trained

[2]The model implementation is available at https://github.com/HPC-HD/ECCO-genre-classification. The original ECCO-Bert model has been released and is available at https://huggingface.co/TurkuNLP/eccobert-base-cased-v1

a linear model by concatenating the tf-idf features of the full text with the pooling output of the fine-tuned ECCO-BERT-Seq. The input can be denoted as $[\Phi_{tfidf}(X_i), \Phi_{ECCO-BERT}(X_i[:510])]$, where Φ represents the transformer encoder and the vectorizer of tf-idf. We call the model ECCO-BERT-tfidf, all results shows in Table 1.

4.2. Baseline Models

There are two baseline models we adopt for comparison. The input of linear model is tf-idf features of the full document. The model only contains the linear layer, the fan-out of the linear model is the number of main or sub categories. The bert-base-cased is released by [11], which we fine-tuned directly with our training data.

5. Results

There are 30,119 documents annotated by experts. 6,024 documents were randomly selected and split into development and test datasets, with 3,012 documents each. The labels contain 10 main categories and 43 sub-categories. The genre labels are presented in A.1.

5.1. Experimental Details

The sequence length of all BERT models is set to be 512. For fine-tuning the ECCO-BERT-Seq model and bert-base-cased model, we only adopt the first 510 sub-words of the document as input. These models are trained for 100 epochs on 1 NVIDIA V100. ECCO-BERT-Chunk is fine-tuned on 4 NVIDIA A100 GPUs; the main category model and the sub-category model were trained for 21 and 20 epoches respectively, using an early stop strategy.

The loss function of the linear model is cross entropy. We perform training for 200 epochs with SGD with momentum [28] and a batch size of 32. The number of tf-idf features is 500,000.

The ECCO-BERT-tfidf models are trained for 220 epochs with SGD with momentum. The feature extractors are the encoders of fine-tuned ECCO-BERT-Seq and vectorizer of linear base models. In order to make the model make more use of tf-idf features, at the first 200 epoches, we mask the features from ECCO-BERT-Seq. The number of tf-idf features is 500,000, the dimension of features extracted from ECCO-BERT-Seq is 768.

In addition to the primary ECCO-BERT model, we also trained the Perceiver IO model [9] on the same data as the BERT models. Perceiver is a Transformer model that decouples input size from overall model size and allows the model to scale linearly with the size of the input as well as model depth. Perceiver IO generalizes Perceiver further by allowing for arbitrary outputs. Due to their linear scaling characteristics, the Perceiver models make it practical to use character-level input data which could result in a model that is more robust against character-level OCR artefacts in the ECCO dataset. Testing this property is our main motivation for using Perceiver IO on this task. We pre-trained Perceiver on the ECCO data for 1 million steps with an effective batch size of 768. Training is done similarly to ECCO-BERT, except that the next sentence prediction task is not used. Fine-tuning for the genre classification task is also similar to the BERT models, except that unfiltered, byte-level data is used as model inputs.

Table 1
Performance comparison for predicting categories with fine-tuned ECCO-BERT models and baselines

	Main categories		Sub categories	
Type	Development (acc)	Test (acc)	Development (acc)	Test (acc)
linear_model	0.9303	0.9333	0.8828	0.8904
bert-base-cased	0.9316	0.9363	0.9011	0.9041
ECCO-Perceiver-Seq	0.9595	0.9555	0.9280	0.9329
ECCO-BERT-Seq	0.9562	0.9602	0.9333	0.9416
ECCO-BERT-Chunk	0.9622	0.9651	0.9346	0.9419
ECCO-BERT-tfidf	0.9645	0.9688	0.9442	0.9485

5.2. Genre Model Performance

We report the models' accuracy for main categories and sub-categories in Table 1. The confusion matrix of ECCO-BERT-tfidf is shown in Figure 2. There is a significant gap between fine-tuned bert-base-cased model and other models based on ECCO-BERT, since the bert-base-cased model is pre-trained on modern language corpus, was not exposed to OCR noise during pre-training, and the language has naturally evolved between 18th century and present-day English. Although ECCO-BERT-Seq is only conditioned on the first 510 tokens of the document, its results are also competitive compared to ECCO-BERT-Chunk and ECCO-BERT-tfidf which consider the full document. As shown in Table 1, ECCO-BERT-tfidf performs best since it combines the transformer feature and tfidf of the full document. ECCO-BERT-tfidf is also much faster than ECCO-BERT-Chunk because extracting tfidf is much faster than inference of transformer models.

Of particular note is the performance of all ECCO-BERT models over base BERT and the linear model, when looking at the more fine-grained categories. Somewhat disappointingly, the fine-tuned Perceiver IO models do not perform better than BERT-based models on this task in our evaluation. This would indicate that the OCR noise does not interfere with the genre detection task enough to degrade the performance of BERT-based models.

5.3. Document-level Evaluation and Prediction results

Here we report on both the evaluation of the document-level results for the main categories. The confusion matrix in 2 shows that the precision of the literature category is the highest while education is the lowest. We also use the ECCO-BERT-tfidf model to predict unlabeled ECCO data and obtain model-predicted genre distributions. There are 177,494 unlabeled documents in total. The breakdown of predicted categories are shown in Figure 3. As our label taxonomy is custom-made, there is no ground truth for the entirety of ECCO to fully evaluate the accuracy of the predictions. However the predictions roughly match up with our expectations: previous analyses of the ESTC, using the existing Dewey Decimal System labels, have found that the most common subject category is religion. [4]

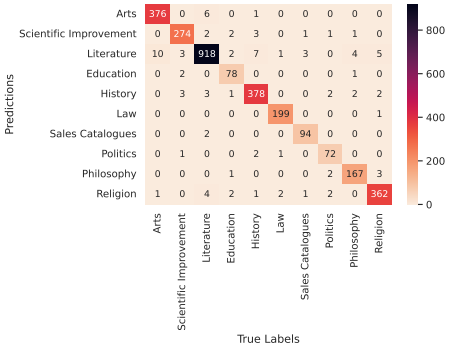

Figure 2: ECCO-BERT-tfidf confusion matrix

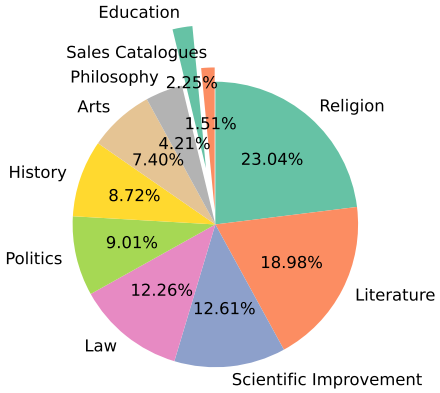

Figure 3: The predicted main categories' distribution.

6. Fine-grained analysis with ECCO-BERT-Seq

6.1. Sequential Genre Change

As well as using the ECCO-BERT-Seq to generate document-level predictions using average values, we can use the individual chunk predictions directly. Here we propose a method to use this paragraph-level detection to detect chunks within documents where the change from one genre to another is significant and sustained. Because the predicted genre generally oscillates significantly from one individual chunk to the next, we needed a method to capture only sustained changes, ignoring shorter breaks within a 'run' of the same genre. To do this, we used the Kleinberg algorithm for detecting 'bursts' of activity in time-series data. [12] This uses a hidden Markov process to probabilistically determine when a subsequent event will occur. When events occur more rapidly and for sustained periods in comparison to this determination, these are labelled bursts. The detection of the bursts were computed using R bursts package [3], which implements the Kleinberg algorithm.

To adapt this method, the most probable prediction for each chunk within each document was treated as a time series data point for Kleinberg. We have calculated sections for main and subcategories separately. The method allows for 'fuzzy' and overlapping sections of genres. Additionally we have experimented with only retaining highly-probable classifications which helped to further filter out noise. There are drawbacks: because the burst method looks for change rather than simply all clusters of events, currently not all sections are detected if most of the text is of a single genre.

To give some examples, we take some exemplary texts and calculate genre bursts. To visualise the changes in genre, top genre predictions (over .5 probability) are charted as a scatterplot in the paragraph sequence, coloured by genre. Burst start and end points are overlaid as coloured areas. As the method looks for periods of change rather than absolute values, it ignores the main category of the book (which is detected by the document-level method successfully anyway) and in most cases highlights sustained excerpts where the detected genre is different to the dominant one. Here, we see that David Hume's *Political Discourses* (Figure 4,

A) contains discrete sections on economics (categorised as scientific improvement), philosophy (a section on the balance of power), history (a section on 'ancient nations') and finally law (a chapter on the idea of the commonwealth). *Wealth of Nations* (Figure 4, B) begins with a section on labour and society categorised here as philosophy and smaller sections on law (a discussion on a specific statute), and in the education genre. Most of the book is not classified as its dominant genre (economics and trade, under the higher-level category scientific improvement) as it does not involve change. Villier's *Miscellaneous Works* (Figure 4, C) detects a large number of overlapping genre changes. Finally, *Robinson Crusoe* (Figure 4, D) is also mostly without detected bursts, but of note is a section of religious genre, corresponding to a section in the plot where Crusoe is ill and has prophetic dreams.

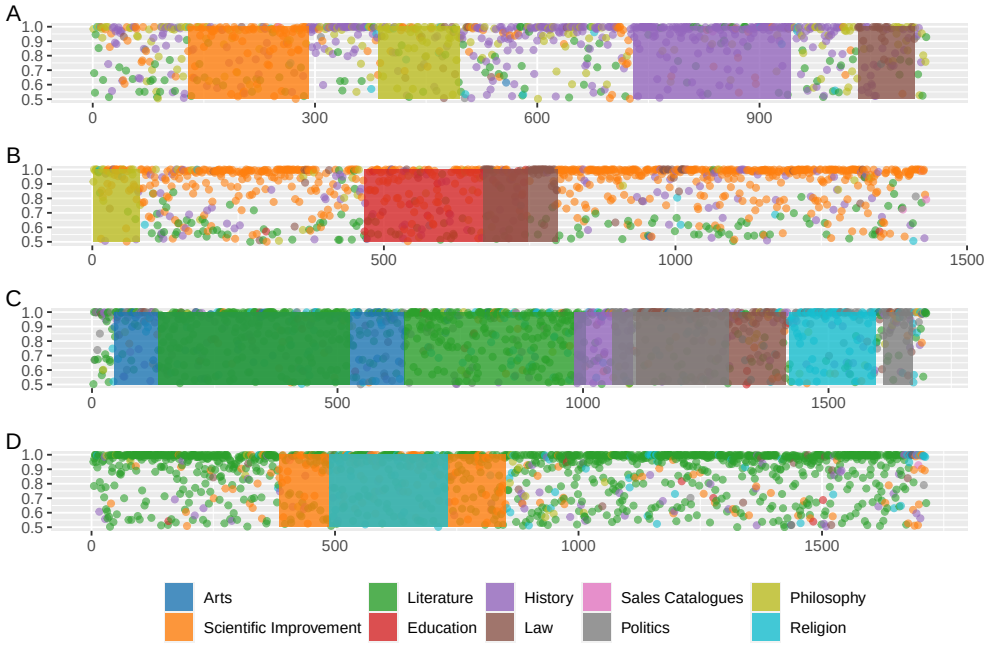

Figure 4: Sections of genre bursts calculated by Kleinberg algorithm. Points are prediction probabilities, filtered to .5 or greater. Coloured shaded areas are sections of genre bursts. Because this method looks for change rather than absolute values, it ignore the main genre of the book in the first two cases (politics, and economics).

7. Discussion and Conclusion

In this paper we aimed to describe the process to detect sections of fuzzy and overlapping genre excerpts within individual editions. The results show that at the level of fine-grained divisions (43 subcategories), a model which combines the tfidf feature of the full document and the features of a fine-tuned ECCO-BERT model performs significantly better than baselines, suggesting they may be particularly useful for such tasks. That the BERT model performed so

well on fine-grained categories is significant because existing methods to look at genre have generally used very broad divisions (such as fiction and non-fiction). The kinds of questions we are interested in use more fine-grained categories, for example looking at the rise of medical textbooks in certain publishers. This kind of sequencing also has other potential uses, for example document retrieval. On the present task, we did not observe any improvement offered by the Perceiver model, which we specifically included to test a character-level model which is capable of accounting for OCR artefacts. At present, we think this is due to a combination of two factors: Firstly, the base performance on the task is around 95% accuracy, leaving only very little headroom for improvement with more advanced models. And secondly, the task is by its nature a document-level task and the good performance of the linear baseline demonstrates that enough information is present in the data even without explicitly accounting for OCR errors. It is therefore possible that the advantages of character-based models such as the Perceiver will be demonstrated on tasks where the correct modelling of individual word occurrences in their context plays a more significant role. These would include various text tagging and information retrieval tasks.

In our future work we hope to further develop the sequencing method, and investigate the genres in their own right, for instance looking at the sequence patterns of individual authors, the relationship between intra-book diversity and the success of particular authors or publishers, and understanding co-occurrence between genres.

References

[1] B. Baptiste, B. Favre, J. Auguste, and C. Henriot. "Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?" In: *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*. 2021.

[2] B. M. Benedict. "The Paradox of the Anthology: Collecting and Différence in Eighteenth-Century Britain". In: *New Literary History* 34.2 (2003), pp. 231–256. URL: http://www.jstor.org/stable/20057778.

[3] J. Binder. *bursts: Markov Model for Bursty Behavior in Streams*. 2022. URL: https://CRAN.R-project.org/package=bursts.

[4] J. Feather. "British Publishing in the Eighteenth Century: a preliminary subject analysis". In: *The Library* s6-VIII.1 (1986), pp. 32–46. DOI: 10.1093/library/s6-VIII.1.32. URL: https://doi.org/10.1093/library/s6-VIII.1.32.

[5] Gale. *Eighteenth Century Collections Online*. URL: https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online.

[6] A. Goyal and V. Prem Prakash. "Statistical and Deep Learning Approaches for Literary Genre Classification". In: *Advances in Data and Information Sciences*. Ed. by S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. Mishra, and B. K. Singh. Vol. 318. Singapore: Springer Singapore, 2022, pp. 297–305. DOI: 10.1007/978-981-16-5689-7_26. URL: https://link.springer.com/10.1007/978-981-16-5689-7%5C%5F26.

[7] S. Gupta, M. Agarwal, and S. Jain. "Automated Genre Classification of Books Using Machine Learning and Natural Language Processing". In: *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. Noida, India: Ieee, 2019, pp. 269–272. DOI: 10.1109/confluence.2019.8776935. URL: https://ieeexplore.ieee.org/document/8776935/.

[8] M. J. Hill and S. Hengchen. "Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study". In: *Digital Scholarship in the Humanities* 34.4 (2019), pp. 825–843. DOI: 10.1093/llc/fqz024. URL: https://academic.oup.com/dsh/article/34/4/825/5476122.

[9] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira. *Perceiver IO: A General Architecture for Structured Inputs & Outputs*. 2021. DOI: 10.48550/arxiv.2107.14795. URL: https://arxiv.org/abs/2107.14795.

[10] M. Jiang, Y. Hu, G. Worthey, R. C. Dubnicek, T. Underwood, and J. S. Downie. "Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts." In: *Chr.* 2021, pp. 266–279.

[11] J. D. M.-W. C. Kenton and L. K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of naacL-HLT*. 2019, pp. 4171–4186.

[12] J. Kleinberg. "Bursty and Hierarchical Structure in Streams". In: *Data Mining and Knowledge Discovery* 7.4 (2003), pp. 373–397. DOI: 10.1023/a:1024940629314. URL: https://doi.org/10.1023/A:1024940629314.

[13] K. Labusch, P. Kulturbesitz, C. Neudecker, and D. Zellhöfer. "BERT for named entity recognition in contemporary and historical German". In: *Proceedings of the 15th conference on natural language processing*. 2019, pp. 9–11.

[14] L. Lahti, E. Mäkelä, and M. Tolonen. "Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming". In: (2020). URL: https://helda.helsinki.fi/handle/10138/327728.

[15] 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. URL: https://doi.org/10.1080/01639374.2018.1543747.

[16] E. Manjavacas and L. Fonteyn. "Adapting vs. Pre-training Language Models for Historical Languages". In: *Journal of Data Mining & Digital Humanities* Nlp4dh (2022). DOI: 10.46298/jdmdh.9152. URL: https://jdmdh.episciences.org/9690.

[17] J. Manson. *A catalogue of the entire and genuine library and prints of Robert Salusbury Gotton, Esq. F.A.S. [electronic resource] : Comprehending an extensive and valuable collection of books of coins, medals and antiquities, with a few fink missals and other manuscripts on vellum, which, with some other select parcels of books lately purchased, are now on sale for ready money, at the price printed in the catalogue, and on the first leaf of each-book, By John Manson, bookseller, No 5, Duke's-Court, St. Martin's-Lane, where catalogues (Price 6d) may be had.* [London, 1789, [2], 102 pages.

[18] D. Mazella, C. Willan, D. Bishop, E. Stravoski, W. Barta, and M. James. ""All the modes of story": Genre and the Gendering of Authorship in the Year 1771". In: *ABO: Interactive Journal for Women in the Arts, 1640-1830* 12.1 (2022). DOI: http://doi.org/10.5038/2157-7129.12.1.1256. URL: https://digitalcommons.usf.edu/abo/vol12/iss1/10.

[19] F. Moretti. *Distant reading*. London ; New York: Verso, 2013.

[20] M. Poovey. "Mary Wollstonecraft: The Gender of Genres in Late Eighteenth-Century England". In: *NOVEL: A Forum on Fiction* 15.2 (1982), pp. 111–126. URL: http://www.jstor.org/stable/1345219.

[21] I. Rastas, Y. C. Ryan, I. L. 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*. The Association for Computational Linguistics. 2022.

[22] J. Raven. *The business of books: booksellers and the English book trade, 1450-1850*. New Haven: Yale University Press, 2007.

[23] I. Rivers, ed. *Books and their readers in eighteenth century England*. Leicester: Leicester Univ. Press [u.a.], 1982.

[24] I. Rivers, ed. *Books and their readers in eighteenth-century England: new essays*. London New York: Leicester University Press, 2001.

[25] T. Schmidt, K. Dennerlein, and C. Wolff. "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language". In: Association for Computational Linguistics. 2021.

[26] T. Schmidt, K. Dennerlein, and C. Wolff. "Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays". In: (2021).

[27] S. Schweter and L. März. "Triple E-Effective Ensembling of Embeddings and Language Models for NER of Historical German." In: *CLEF (Working notes)*. 2020.

[28] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. Pmlr. 2013, pp. 1139–1147.

[29] K. Todorov and G. Colavizza. "An Assessment of the Impact of OCR Noise on Language Models". In: *arXiv preprint arXiv:2202.00470* (2022).

[30] M. Tolonen, E. Mäkelä, A. Ijaz, and L. Lahti. "Corpus Linguistics and Eighteenth Century Collections Online (ECCO)". In: *Research in Corpus Linguistics* 9.1 (2021), pp. 19–34. DOI: 10.32714/ricl.09.01.03. URL: https://ricl.aelinco.es/index.php/ricl/article/view/161.

[31] M. Tolonen, E. Mäkelä, A. Ijaz, and L. Lahti. "Corpus Linguistics and Eighteenth Century Collections Online (ECCO)". In: *Research in Corpus Linguistics* 9.1 (2021), pp. 19–34. DOI: 10.32714/ricl.09.01.03. URL: https://ricl.aelinco.es/index.php/ricl/article/view/161.

[32] 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.

[33] T. Underwood. *Distant horizons: digital evidence and literary change*. Chicago: The University of Chicago Press, 2019.

[34] T. Underwood. "Genre Theory and Historicism". In: *Journal of Cultural Analytics* 2.2 (2016). DOI: 10.22148/16.008. URL: https://culturalanalytics.org/article/11063.

[35] T. Underwood. "The Life Cycles of Genres". In: *Journal of Cultural Analytics* 2.2 (2016). DOI: 10.22148/16.005. URL: https://culturalanalytics.org/article/11061.

[36] T. Underwood. "Understanding Genre in a Collection of a Million Volumes, Interim Report". In: (2014). DOI: 10.6084/m9.figshare.1281251.v1. URL: https://figshare.com/articles/journal%5C%5Fcontribution/Understanding%5C%5FGenre%5C%5Fin%5C%5Fa%5C%5FCollection%5C%5Fof%5C%5Fa%5C%5FMillion%5C%5FVolumes%5C%5FInterim%5C%5FReport/1281251.

[37] T. Underwood, M. L. Black, L. Auvil, and B. Capitanu. *Mapping Mutable Genres in Structurally Complex Volumes*. 2013. DOI: 10.1109/BigData.2013.6691676. URL: http://arxiv.org/abs/1309.3323.

[38] J. Worsham and J. Kalita. "Genre Identification and the Compositional Effect of Genre in Literature". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1963–1973. URL: https://aclanthology.org/C18-1167.

[39] H. Yoo, J. Jin, J. Son, J. Bak, K. Cho, and A. Oh. "HUE: Pretrained Model and Dataset for Understanding Hanja Documents of Ancient Korea". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 1832–1844. URL: https://aclanthology.org/2022.findings-naacl.140.

A. Appendix

A.1. The main categories and sub-categories

Table 2
The information of main categories and sub-categories

Main categories	Sub-categories
Arts	Fine Arts and Aesthetics
	Music, hymns, songs
	Theatre, plays, opera
Education	Advice literature
	General Education
	Recipe Books
	Hobbies & Games
	Instructional books
History	Biographical History
	General History
Law	Acts, proclamations
	Appeals
	Collected bills
	Legal essays
	Proclamations
	Trial accounts
Literature	Classics
	Collected Works
	Criticism
	Drama
	Novels
	Other fiction
	Periodicals
	Poetry
	Travel
Philosophy	Human understanding, metaphysics
	Moral Philosophy
	Political philosophy
Politics	Political essays
	Intelligence
	Parliamentary speeches
Religion	Sermons
	Theology
Sales Catalogues	Sales catalogues, almanacs, directories etcetera
Scientific Improvement	Agriculture and animal husbandry
	Economics and trade
	Geography, cartography, astronomy and navigation
	Languages
	Mathematics
	Medicine and anatomy
	Natural history
	Natural philosophy
	Practical trades, mechanics, engineering