In Search of Lost Adventure Novels: Supervised Genre Retrieval and Corpus Refinement in Gallica

Jean Barré
Lattice, ENS-PSL, Paris, France
jean.barre@ens.psl.eu

Abstract

This paper addresses a practical problem in computational literary history: retrieving adventure novels (*roman d'aventures*) from a large digitized collection of French fiction where genre metadata are sparse and unreliable. We begin with supervised genre modeling based on a historically situated seed list of 101 adventure novels drawn from literary scholarship. We compare several classifiers and representations, and validate them against 364 independently labeled adventure novels from the *Chapitres* corpus. The best-performing model, HistGradientBoosting on mean paragraph embeddings, achieves strong external recall (81%) despite the small training set. We then apply this model to the 12,176-novel *Fictions littéraires de Gallica* archive and refine the resulting candidate corpus through a graph-based post-processing step over a k-nearest-neighbor similarity graph. On the *Chapitres* benchmark, this graph correction produces negligible changes in retrieval performance, indicating that it is not a generally superior classifier. On Gallica, however, it yields a more cohesive and homogeneous candidate corpus and surfaces interpretable correction cases, including missed canonical adventure novels and excluded borderline texts. We therefore argue that graph-based correction is best understood not as a replacement for supervised classification, but as a heuristic for refining large, noisy archival corpora where exhaustive manual annotation is impossible.

1 Introduction

Genre fiction occupies a singular place in the circulation and consumption of narrative prose over the last two centuries. These novels were not marginal in their own time: they were mass readings, editorial successes, and shared cultural references that helped shape the imaginative horizons of generations of readers. Yet a large part of this production now belongs to what Margaret Cohen calls literature "*hors d'usage*": texts that are no longer read and rarely studied, despite their historical visibility and commercial centrality (Cohen, 1999). Literary history has often conflated "literature" with a narrow fraction of production stabilized by consecration, schooling, and transmission.

Modern literary study typically works on a tiny sample relative to everything that was published and consumed: a *canon* that represents only a minute part of a vast continent of works (Moretti, 2000a,b). The methodological and epistemological stake is not simply to add forgotten works to an already constituted corpus, but to shift the focal length: to open investigation to the massive *archive* of printed production that exceeds the canon by orders of magnitude. For popular and serial fiction, this shift is especially important, because the forms that mattered most historically were often not the ones best preserved by critical memory.

This canon/archive opposition provides a decisive framework for the study of popular literature. As Matthieu Letourneux has shown, much of genre fiction belongs to what he calls "factory fiction" (*fiction à la chaîne*): works embedded in industrial, serial, and media dynamics where meaning and value are partly constructed at the level of aggregates (series, collections, formats) rather than the single work (Letourneux, 2017). Genre fiction is thus a privileged observatory for understanding how narrative forms stabilize, circulate, and transform under the constraints of editorial production and reader expectations.

But moving from the canon to the archive immediately raises a methodological problem: how can we produce knowledge about thousands of novels without reproducing, under another name, the restrictive logic of the canon? This is the ambition of *distant reading* (Moretti, 2013), and more broadly the field of Computational Literary Studies (Allison et al., 2011; Jockers, 2013). In practice, however, archive-scale literary history depends on a prior technical task that is often underdescribed: build-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 255–263

ing the corpus itself. If metadata are incomplete, unreliable, or too coarse-grained, retrieving a historically meaningful set of texts becomes a problem of inference rather than simple catalog search.

In this paper, we adopt that program for a concrete and difficult case: the French adventure novel (*roman d'aventures*). The genre flourished from the 1870s to the 1930s and played a central role in popular print culture (Letourneux, 2010). It is also, in many respects, *the archive par excellence*: an industrial genre that was massively produced, widely read, and later largely forgotten, leaving behind a long tail of minor names and ephemeral titles. Its boundaries are porous, its definition debated, and it hybridizes easily with neighboring forms (historical fiction, exotic fiction, maritime fiction, feuilleton drama). From the point of view of corpus construction, it is therefore an ideal stress test: historically central, bibliographically unstable, and difficult to recover by metadata alone.

Our approach combines supervised genre modeling with archive-scale corpus refinement. First, we train several classifiers on a historically anchored seed list: the adventure novels cited in Letourneux's literary history of the genre, matched with negatives drawn from the same digitized collection. This follows what Underwood (2019) calls "perspective modeling": the goal is not to define the genre once and for all, but to operationalize a situated critical perspective and observe which textual regularities correlate with it. We evaluate these models against an independent benchmark, the *Chapitres* corpus, in order to select the most robust configuration for large-scale retrieval.

Second, once the best-performing supervised model has been chosen, we apply it to the full Gallica archive and refine its output using the relational structure of the corpus itself: a k-nearest-neighbor similarity graph built over all novels. Texts predicted as adventure but isolated among non-adventure neighbors are treated as likely outliers; texts not initially retrieved but surrounded by predicted-adventure neighbors are treated as likely infiltrators.

This second stage uses local neighborhood agreement in embedding space as a heuristic proxy for relational compatibility among candidate adventure novels. As we show, this relational correction does not improve benchmark retrieval on the smaller validation corpus, but it proves useful as a corpus-refinement heuristic at archive scale, where noise, sparsity, and boundary cases are more pronounced.

The contribution of the paper is therefore threefold. First, it shows that a small, historically situated seed list can support robust retrieval of adventure novels on an independent scholarly benchmark. Second, it clarifies the methodological status of graph-based post-processing: not a better classifier in general, but a useful heuristic for reorganizing candidate corpora in large unlabeled archives. Third, it argues that corpus construction itself can be a site of literary-historical insight, because the cases added or removed by the refinement step expose precisely the borderline zones where genre theory expects ambiguity to reside.

2 Related Work

2.1 Computational approaches to genre

Automatic genre classification has a long history in corpus linguistics, information science, and computational stylistics. Early work (Biber, 1988) in register and genre variation showed that textual categories are rarely defined by a single marker; rather, they manifest as bundles of co-occurring features. Genre detection systems combined surface indicators (character n-grams, function-word frequencies, part-of-speech distributions) with supervised learning, demonstrating robust genre discrimination (Kessler et al., 1997).

Within computational literary studies, the large-scale availability of digitized fiction renewed interest in genre as a measurable regularity. Cranenburgh et al. (2024) combine topic modeling with Biber-style linguistic features on more than 9,800 novels and report subgenre prediction around 70% accuracy. Hettinger et al. (2015, 2016) compared several classifiers on nearly 1,700 German novels and found that linear SVMs with topic-model features achieved the best results. For French, Schöch (2021) showed that topic modeling could distinguish dramatic subgenres in classical and Enlightenment theater.

Recent work has expanded the methodological repertoire of computational literary studies toward neural encoders, transformer-based models, and large language models. Hatzel et al. (2023) show that these methods are now prominent in CLS, while traditional feature-based approaches remain important because of their transparency and fit with domain-specific literary questions. Bamman et al. (2024) frame classification in cultural analytics as a sensemaking practice rather than mere automation: classifiers can test categories —such as genre

labels— and challenge their boundaries. Closer to the present study, Barré (2024) proposed a passage-level operationalization of adventure architextuality by detecting scenes of danger and exploration with LLMs. The present paper shifts from local generic scenes to document-level retrieval: identifying adventure novels as archival objects in a large, weakly labeled corpus.

2.2 Perspectivist modeling and historically situated labels

A widely influential response to label uncertainty is Underwood's *perspectivist modeling*. The core proposal is to treat genres not as timeless essences but as historically situated categories embedded in institutional practices. Training lists may come from library catalogs, editorial series, or period-specific critical corpora; models trained on these lists aim to reproduce a given classificatory perspective (Underwood, 2019, 2016). Perspectivist modeling also clarifies what model errors mean: false positives and false negatives are not merely failures to be minimized; they are probes into boundary cases, hybrid zones, and mismatches between textual regularity and institutional labeling.

This perspective is particularly relevant for archival retrieval. When the aim is not to optimize a benchmark but to reconstruct a historically meaningful corpus from sparse metadata, the question shifts from "what is the true genre label?" to "what literary-historical viewpoint is being operationalized, and what does it recover?"

2.3 Genre as family resemblance

Genre theory has long insisted that membership is graded and boundaries porous. For popular fiction, Letourneux (2017) emphasizes how conventions circulate through serial production, editorial packaging, and interdiscursive context. This view resonates with family-resemblance accounts of categorization (Wittgenstein, 1958): a text can be *more or less* adventure depending on how densely it shares motifs, rhetoric, and narrative economy with other adventure texts.

In computational terms, this suggests that genre may be modeled not only as a property of isolated documents but also as a position in a similarity space. Recent work has pushed this relational perspective by adopting variable-granularity approaches: Calvo Tello (2021) and Henny-Krahmer (2023) apply multi-label classification to textual segments, yielding networks in which edges reflect generic proximity at different scales. Our graph-based refinement step belongs to this broader line of thought, though in a deliberately simple form: we use local neighborhood structure not to replace supervised classification, but to test whether the retrieved corpus becomes more coherent when generic membership is treated relationally. We return in Section 4.3 to the conceptual gap between family resemblance and geometric proximity, which is not a logical entailment but a working postulate of our pipeline.

3 Data and Representation

3.1 Base corpus

We start from the *Fictions littéraires de Gallica* collection (19,240 digitized novels) (Langlais, 2021). Following Barré (2024), we apply preprocessing filters: we keep texts with estimated OCR quality above 95%, exclude complete works volumes, and remove re-editions. After deduplication and filtering, the working corpus contains 12,176 novels.

This filtering matters because adventure fiction is especially vulnerable to archival noise. Popular novels are frequently reprinted, retitled, or embedded in heterogeneous volumes, and OCR noise can disproportionately affect precisely the low-status, weakly curated material that is most relevant to archive-scale literary history. Our goal is not to recover *all* instances of adventure fiction in Gallica, but to construct a stable working corpus from the subset of novels whose textual signal is strong enough to support reliable modeling.

3.2 Text representation

Each novel is represented as a vector of Most Frequent Words (MFW) frequencies, a standard stylometric representation. We use 5,000 MFW, selected empirically by comparing balanced accuracy as the feature count increases; performance improves up to 5,000 and then plateaus.

We also rely on paragraph-level vector representations produced by a fine-tuned multilingual embedding model (based on BGE-M3) specialized for French literary texts (Barré, 2024). Because novels exceed the model's context window, we encode the text as a sequence of paragraph embeddings and compute the mean to obtain a single fixed-size representation per novel. This aggregation preserves a global semantic/stylistic signal while remaining compatible with standard classifiers and similarity-graph construction at corpus scale.

The two representations serve different strengths. BoW vectors are transparent and robust on small datasets, while mean paragraph embeddings offer a richer notion of proximity that is especially useful when comparing novels across a heterogeneous archive. Rather than assuming in advance which representation is best, we evaluate both on the independent validation corpus and use those results to guide model selection.

4 Method

4.1 Training data

Following a perspectivist modeling logic, we construct a positive seed list from Letourneux's (2010) literary history of the adventure novel: 101 cited adventure novels present in the Gallica corpus. We pair this initial set with negatives sampled from the same collection.

The Letourneux list is methodologically convenient because it is diverse and reduces authorial idiolect bias: it spans 36 distinct authors, with at most five novels per author. To make the negative class more representative of the archive, we broaden it beyond a simple one-to-one match and sample non-adventure novels across the temporal range of the corpus, including up to five titles per decade outside the genre's core production window (1650–1870 and 1930–1950). The final training set contains 205 novels.

This design has two advantages. First, it keeps the positive class historically meaningful: we are not training on a diffuse modern notion of adventure, but on a literary-historical list produced for a different scholarly purpose. Second, and crucially for the validity of the classifier, the negative class spans both the same decades as the positives (1870–1930) and earlier and later periods. A model that distinguished positives from negatives on the basis of publication date alone would therefore perform poorly on this training set; the strong external recall on *Chapitres* novels (Section 5.1), which span an even broader temporal range, provides additional indirect evidence that the model has not collapsed to a period detector. A more targeted ablation with strictly date-matched negatives would further sharpen this point and is left for future work.

4.2 Supervised models

We compare three classifiers (linear SVM (Cortes and Vapnik, 1995), Random Forest, and HistGradientBoosting), using SCIKIT-LEARN (Pedregosa et al., 2011). Each is tested on two representations: BoW vectors based on 5,000 Most Frequent Words, and mean paragraph embeddings derived from the BGE-M3-based model fine-tuned on French literary texts (Chen et al., 2024; Barré, 2024).

We evaluate models in two stages. First, we report 5-fold cross-validation on the Letourneux-based training set, using stratified splits at the novel level. We do not impose group constraints by author, since the seed list is already designed to limit author bias (at most five novels per author, distributed across 36 distinct authors in the positive class). Second, and more importantly, we test each model on the independently labeled *Chapitres* corpus. *Chapitres* is not the target corpus of the paper, but a methodological bridge between a small perspectivist training set and a much larger unlabeled archive: it allows us to select the most robust model before deployment on Gallica.

4.3 Graph-based correction

Once a supervised model has been selected, we refine its predictions through a graph-based post-processing step. The intuition is relational: if adventure novels share conventions, a randomly chosen adventure novel should have a substantially higher proportion of co-labeled neighbors in its k nearest neighbors than would be expected by chance. Throughout this paper, we use "neighborhood" in this restricted sense — a property of a node's k-NN list — rather than as a claim about absolute density in embedding space. The k-NN graph alone is, by construction, blind to how the corpus occupies the space: every node has exactly k outgoing neighbors regardless of its position. The pairwise similarity diagnostics below complement the graph view by characterizing actual proximity.

The refinement step rests on a working postulate that we want to make explicit: *family resemblance among adventure novels is at least partially expressed by proximity in our embedding space*. This correspondence is not a logical entailment — family resemblance is a structural notion that does not, in general, require metric structure — and we do not establish it formally. We treat it as a working hypothesis indirectly supported by two orthogonal observations: (i) supervised classifiers trained on Letourneux generalize to the independently annotated *Chapitres* adventure set with recall above 0.78 (Section 5.1), indicating that the embedding space encodes a substantial portion of the genre signal recognized by external annotators; and (ii) the tem-

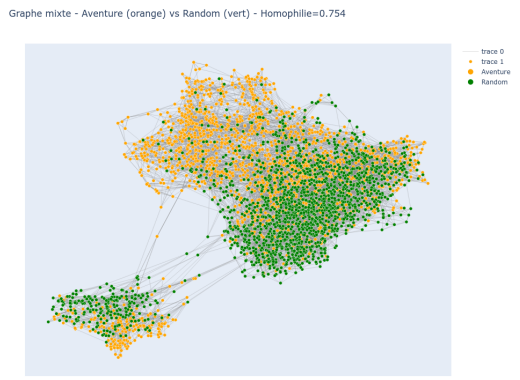

Figure 1: k-NN similarity graph: adventure candidates retrieved by the supervised model (orange) vs. randomly sampled non-adventure novels (green). The visible label separation motivates neighborhood-based refinement.

poral distribution of retrieved candidates on Gallica converges with Letourneux's independently established periodization of the genre (Section 5.3). Neither establishes the postulate, but together they motivate using neighborhood structure as an additional, relational source of evidence for membership.

We then apply a simple neighborhood rule on a directed k-NN graph built over the corpus using cosine similarity on mean paragraph embeddings:

Outlier removal: a novel predicted as adventure is removed if at least 80% of its k nearest neighbors are predicted as non-adventure.

Infiltrator addition: a novel not predicted as adventure is added if at least 80% of its k nearest neighbors are predicted as adventure.

This correction is deliberately simple and interpretable. Its value must be evaluated differently depending on the corpus: by external retrieval metrics when independent labels are available, and by internal cohesion diagnostics plus qualitative inspection when they are not.

5 Validation and Archive-Scale Retrieval

5.1 External validation on *Chapitres*

To assess whether the models retrieve genuine adventure novels beyond their own training signal, we evaluate them against an independent source of genre labels: the *Chapitres* corpus (Leblond, 2022), a collection of approximately 2,960 French novels annotated by literary scholars. Among these, 364 are tagged as *romans d'aventures*. The comparison is deliberately conservative: the Letourneux seed list reflects a literary-historical perspective centered on the 1870–1930 period, whereas *Chapitres* labels were assigned independently and may reflect partially different genre conceptions.

Classifier	Repr.	BAcc	F1	AUC
SVM linear	BoW	.87±.04	.87±.04	.94±.03
RandomForest	BoW	.88±.06	.88±.06	.95±.03
HistGradBoost	BoW	.85±.05	.85±.05	.94±.02
SVM linear	Emb	.86±.05	.85±.06	.93±.02
RandomForest	Emb	.89±.05	.89±.05	.95±.03
HistGradBoost	Emb	.88±.04	.88±.04	.95±.03

Table 1: Cross-validation (5-fold, stratified random splits at the novel level) on the Letourneux training set (n=205). BoW = 5,000 MFW; Emb = mean paragraph embeddings (1,024-d).

Classifier	Repr.	P	R	F1	Adv.
SVM linear	BoW	.42	.83	.56	302/364
RandomForest	BoW	.43	.82	.57	298/364
HistGradBoost	BoW	.42	.84	.56	305/364
SVM linear	Emb	.48	.78	.59	283/364
RandomForest	Emb	.48	.84	.61	307/364
HistGradBoost	Emb	.52	.81	.63	294/364

Table 2: External validation on *Chapitres*. "Adv." = adventure novels correctly retrieved among the 364 tagged adventure novels.

Table 1 reports 5-fold cross-validation on the Letourneux training set. Table 2 reports external validation on *Chapitres*. Across all tested models, recall remains high (78–84%), confirming that the Letourneux-anchored prototype generalizes well to an independently labeled corpus.

The best overall external performance is obtained by HistGradientBoosting on mean paragraph embeddings, which achieves the highest F1 (0.634), the highest precision (0.521), and a strong recall (0.808). We therefore select this model for large-scale prediction on Gallica.

The modest precision values do not straightforwardly indicate model failure. Because only part of the *Chapitres* corpus received genre labels, many predicted false positives are plausibly adventure-like or borderline cases. In an archival setting, the absence of a genre label is not equivalent to the absence of genre traits — it often reflects the practical limits of manual annotation. A larger version of this work should include a targeted manual audit of predicted positives not labeled as adventure in *Chapitres*, in order to distinguish genuine false positives from adventure-adjacent or partially annotated cases.

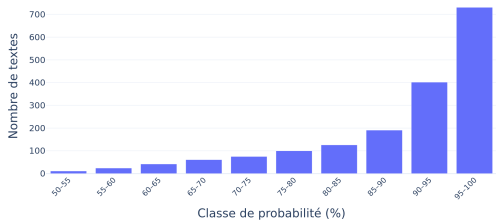

Figure 2: Distribution of predicted adventure probabilities across the full Gallica corpus.

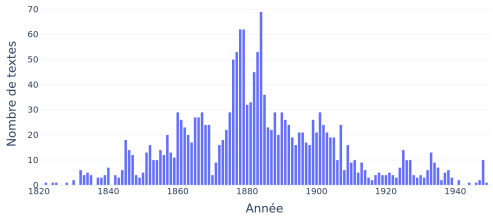

Figure 3: Temporal distribution of supervised adventure predictions on Gallica.

5.2 Graph correction on *Chapitres*

We next test whether graph-based post-processing improves retrieval on the independent benchmark. Applying the same correction rule to the HistGradientBoosting+Emb predictions on *Chapitres* produces only negligible changes. In the configuration retained for the Gallica experiments (k=10, τ=80%), the correction removes 17 outliers and adds 18 infiltrators, but F1 moves from 0.634 to 0.631, and recall remains stable. We tested several configurations ($k \in \{5, 10, 15, 20\}$ and $\tau \in \{70\%, 80\%, 90\%\}$), and none improved on the supervised baseline.

This negative result is important: graph-based correction is not a generally superior classifier. On a smaller, relatively well-annotated corpus, the supervised model already captures most of the available signal, leaving little room for neighborhood-based correction to add value. At the same time, this null result clarifies the status of the second stage: if graph refinement is useful, it will be useful where metadata are sparse, candidate sets are larger, and systematic errors have more room to accumulate. That is precisely the situation on Gallica.

5.3 Prediction on the Gallica archive

Having selected HistGradientBoosting+Emb on the basis of external validation, we apply it to the full *Fictions littéraires de Gallica* corpus (12,176 novels). The supervised model retrieves an initial set of candidate adventure novels, which we then refine with the same graph-based rule.

Because Gallica lacks exhaustive independent labels, evidence on this corpus is necessarily indirect. We assess the graph refinement step not through external retrieval metrics, but through internal corpus-structure diagnostics and qualitative inspection.

The supervised predictions are not randomly distributed. Figure 2 and Figure 3 show that they concentrate in the late nineteenth and early twentieth centuries, with a peak broadly consistent with Letourneux's periodization of the genre, and the model assigns high confidence to a large fraction of the retrieved candidates. These diagnostics do not prove correctness, but they suggest that the model is recovering a historically plausible region of the archive rather than diffusing predictions uniformly across the corpus.

Metric	Before	After	Δ
Novels	1,375	1,391	+16
Mean sim.	0.789	0.825	+3.6%
Std. dev.	0.098	0.080	−18.4%
Homophily	0.754	0.795	+5.4%

Table 3: Effect of graph-based refinement on the Gallica candidate corpus (k=10, τ=80%).

5.4 Graph-based refinement on Gallica

On Gallica, graph correction removes 113 outliers and adds 129 infiltrators, yielding a final corpus of 1,391 adventure novels (Table 3). The refined corpus is more cohesive (+3.6% mean similarity), more homogeneous (−18.4% standard deviation), and more locally clustered (+5.4% homophily) than the supervised output alone.

These gains do not establish that graph refinement is more accurate in an absolute sense — Gallica lacks exhaustive independent genre labels. What they show is that the refined corpus is structurally more compact and internally more consistent.

We acknowledge a partial circularity in this evaluation: the cohesion metrics (mean similarity, homophily) are computed on the same embedding space used to construct the k-NN graph, so part of the observed improvement is mechanically expected from any procedure that relabels in the direction of local label homogeneity. Less circular evidence comes from three sources: (i) the quali-

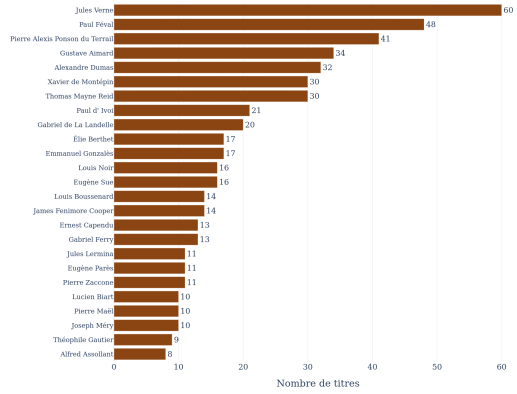

Figure 4: Top 25 authors by title count in the refined adventure corpus (1,391 texts).

tative correction cases (Section 5.5), which can be inspected against external knowledge of the genre; (ii) the temporal convergence of the retrieved corpus with Letourneux's independently established periodization, which is computed outside the embedding space; and (iii) the contrast with the null result on *Chapitres* (Section 5.2), which shows that the same procedure does not automatically improve external retrieval metrics when independent labels are available.

On the *Chapitres* benchmark corpus, graph correction altered only a few dozen labels and yielded no measurable performance gain. On Gallica, it reclassifies 242 novels and visibly changes corpus structure — suggesting that the procedure is sensitive to archival conditions rather than to genre signal alone.

5.5 Corpus structure and qualitative correction cases

The refined corpus of 1,391 adventure novels is anchored by canonical producers such as Verne (60 titles), Féval (48), and Ponson du Terrail (41), but structurally dominated by a long tail: 68.5% of authors appear with a single title, yet collectively account for 27.6% of the corpus (Figure 4). Archive-scale retrieval complicates a canon-centered view of the genre: the adventure novel is not only a handful of famous names, but also a dense background of minor and forgotten production.

The graph-based correction produces diagnostically meaningful cases that clarify what the supervised model learns, and what it misses. Because the graph evaluates each prediction relative to its local neighborhood, it tends to surface structural errors: texts that may appear adventurous in isolation (lexical cues, thematic motifs) but have very few co-labeled neighbors and, conversely, texts that the classifier misses despite being surrounded by predicted-adventure neighbors.

Outliers removed (false positives). We remove 113 predicted-adventure novels flagged as outliers. The removals fall into recurring categories. First, texts clearly outside the adventure economy: moral tales and domestic fiction, such as Girardin's *La famille Gaudry* (1909), may share lexical cues with adventure but have neighborhoods dominated by non-adventure predictions. Second, parodies and humorous texts: Alphonse Allais's *Le Captain Cap*, a sustained parody of adventure fiction, reproduces the vocabulary and topoi of the genre while subverting its narrative economy. The supervised model captures the lexicon and motifs of the genre, but not whether those conventions are being used seriously, ironically, or parodically. This is a constitutive limitation of BoW-based and embedding-based approaches: they detect the availability of generic material, not the pragmatic attitude of the author toward it. We note that a quantitative comparison of the rate of parodic and ironic texts among outliers vs. retained candidates would convert this reading from suggestive to systematic; we leave this targeted study for future work.

Infiltrators added (false negatives). We add 129 infiltrators. These recover canonical adventure works initially missed by the classifier: Jules Verne's *Le Chancellor* (1878), *Famille sans nom* (1889), *Michel Strogoff* (1876), and *L'Étoile du Sud* (1884) — all undisputed classics whose absence from the initial retrieval constituted a notable gap. The correction also recovers central subgenre producers whose stylistic surface diverges from the learned prototype, such as Gustave Aimard, a prolific author of colonial and frontier adventure, and Mayne Reid, a foundational figure for the maritime and frontier subgenre, which initial supervised classification under-recognized.

A high proportion of adventure-labeled k-nearest neighbors functions as an indicator of generic proximity: even when the supervised model under-recognizes a text because of lexical or stylistic divergence, the k-NN graph places it among many adventure-like neighbors. Graph refinement thus functions less as a benchmark-optimizing device than as a way of surfacing missed canonical texts in a large archival corpus. The full list of 113 outliers removed and 129 infiltrators added is provided in

the companion repository (see Limitations).

6 Discussion

Historically situated supervised modeling is robust. The main result of the paper is that a small, historically grounded training list can support robust genre retrieval at scale. Despite the limited size of the Letourneux seed set, all tested models retrieve a large majority of the adventure novels independently labeled in *Chapitres*. This confirms that perspectivist modeling can operationalize a literary-historical category without requiring an exhaustive definition of genre. Strong archive-scale retrieval does not require massive labeled datasets, provided the training signal is coherent and historically grounded.

The best model is embedding-based. External validation clarifies model selection. While BoW and embedding representations both perform well, the best results are obtained by HistGradientBoosting on mean paragraph embeddings. For the large-scale Gallica experiment, we therefore privilege the best externally validated model rather than the most interpretable one. This choice strengthens the retrieval pipeline, even if interpretability must then rely more on downstream qualitative analysis than on linear coefficients alone. It also suggests that adventure fiction is not reducible to a narrow lexical profile: the richer semantic-stylistic signal captured by embeddings transfers better across corpora.

Graph refinement is useful at archive scale, but not as a benchmark gain. The most informative contrast concerns the graph-based correction step. On *Chapitres*, it alters only a small number of labels and does not improve retrieval metrics under any tested configuration. On Gallica, it produces substantial reclassification and yields a corpus that is more cohesive and more homogeneous according to internal structural diagnostics. The graph step should not be understood as a universally better classifier, but as a heuristic for refining candidate corpora in large, noisy archives where exhaustive annotation is unavailable.

In benchmark settings, success means agreement with annotated labels. In large historical archives, the problem is prior to benchmarking: one must first construct a usable corpus from sparse metadata and texts that no team of scholars could annotate exhaustively. Retrieval pipelines should therefore be evaluated not only by classification metrics, but also by their capacity to produce interpretable and historically plausible candidate corpora.

Implications for corpus construction. Archive-scale retrieval reveals the weight of the long tail. For popular genres, this long tail is not peripheral noise; it is one of the main sites where conventions stabilize through repetition, variation, and industrial circulation. The pipeline proposed here is designed for that archival reality: supervised modeling provides a historically anchored first pass, while graph-based refinement helps reorganize the retrieved corpus into a more coherent candidate set for further literary-historical analysis. Corpus construction is not merely preparatory work before interpretation; it is itself part of interpretation, because the operations that define the corpus also expose the structure and ambiguity of the genre being studied.

Limitations

Our approach is constrained both by representation and by the epistemology of labels. Mean paragraph embeddings provide a strong global signal, but they do not directly model plot dynamics, scene sequencing, or narrative role structure. Borderline genres and parodic texts may therefore remain difficult to distinguish. The *Chapitres* validation corpus is independent and valuable, but it is not exhaustive; precision scores likely reflect both genuine overprediction and partial annotation. Conversely, Gallica lacks external labels altogether, so evidence there remains indirect: graph-based refinement can be assessed through internal cohesion diagnostics and qualitative inspection, but not through benchmark retrieval accuracy. Our neighborhood rule (k=10, τ=80%) is also deliberately simple and interpretable; future work could explore softer forms of propagation, weighted neighborhoods, or community-aware refinement procedures.

More targeted analyses would strengthen specific claims of the paper but exceed the scope of a short retrieval contribution. A manual audit of a stratified sample of false positives on *Chapitres* would convert the partial-annotation argument from plausible to demonstrated. A quantitative comparison of the rate of parody, irony, and pragmatic-stance shifts among outliers vs. retained candidates would convert the modal-stance reading of cases like *Le Captain Cap* into a systematic finding. A stricter date-matched negative ablation would provide an additional robustness check against residual period effects.

References

Sarah Allison, Mark Algee-Hewitt, Ryan R. Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. 2011. Quantitative formalism: an experiment. *Pamphlets of the Stanford Literary Lab*, 1(1).

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. *Preprint*, arXiv:2410.12029.

Jean Barré. 2024. Détection automatique de l'architextualité dans le roman d'aventures. In *Humanistica 2024*, Stylométrie, Meknès, Morocco. Association francophone des humanités numériques.

Jean Barré. 2024. Latent structures of intertextuality in french fiction. In *Proceedings of the Computational Humanities Research Conference 2024*, pages 21–26.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

José Calvo Tello. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arxiv:2402.03216 [cs].

Margaret Cohen. 1999. *The sentimental education of the novel*. Princeton University press, Princeton.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Andreas van Cranenburgh, Laura Allen, Serge Sharoff, and Karina van Dalen-Oskam. 2024. Computational methods for the analysis of fiction genres. In *Multidisciplinary Views on Discourse Genre*. Routledge.

Hans Ole Hatzel, Haimo Stiemer, Chris Biemann, and Evelyn Gius. 2023. Machine learning in computational literary studies. *it – Information Technology*, 65(4-5):200–217.

Ulrike Henny-Krahmer. 2023. *Genre Analysis and Corpus Design: Nineteenth-Century Spanish-American Novels (1830–1910)*. Number 17 in SIDE. IDE.

Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. Genre classification on German novels. In *Proceedings of the 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 249–253, Valencia, Spain. IEEE.

Lena Hettinger, Fotis Jannidis, Isabella Reger, and Andreas Hotho. 2016. Significance Testing for the Classification of Literary Subgenres. In *DH2016 book of abstracts*.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*, 1 edition. University of Illinois Press.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schuetze. 1997. Automatic Detection of Text Genre. *arXiv preprint*.

Pierre-Carl Langlais. 2021. Fictions littéraires de gallica.

Aude Leblond. 2022. Corpus chapitres.

Matthieu Letourneux. 2010. *Le roman d'aventures: 1870-1930*. Presses Universitaires de Limoges et du Limousin.

Matthieu Letourneux. 2017. *Fictions à la chaîne: littératures sérielles et culture médiatique*. Poétique. Éditions du Seuil, Paris.

Franco Moretti. 2000a. Conjectures on World Literature. *New Left Review*, 1(1):54–68.

Franco Moretti. 2000b. The Slaughterhouse of Literature. *Modern Language Quarterly*, 61(1):207–228.

Franco Moretti. 2013. *Distant reading*. Verso.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Christof Schöch. 2021. Topic modeling genre: An exploration of french classical and enlightenment drama. *Preprint*, arXiv:2103.13019.

Ted Underwood. 2016. The Life Cycles of Genres. *Journal of Cultural Analytics*, 2(2).

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. The University of Chicago Press.

Ludwig Wittgenstein. 1958. *Philosophical Investigations*, 2 edition. Basil Blackwell.

A Full correction lists

The companion repository provides the complete lists of outliers removed (n=113) and infiltrators added (n=129) during graph-based refinement on the Gallica corpus, together with their nearest-neighbor profiles and supervised-model probabilities. The full retrieved adventure-novel corpus (1,391 titles), with per-novel scores from both the supervised model and the graph-based refinement step, is also made available in the same repository: `github.com/crazyjeannot/In-Search-for-Lost-Adventure-Novels`.