

In Search of Lost Adventure Novels

Supervised Genre Retrieval and Corpus Refinement in Gallica

Jean Barré

NLP4DH – ACL 2026

École Normale Supérieure – PSL & LaTTiCe (CNRS, Sorbonne Nouvelle)

What This Talk Is About

The problem

How do we retrieve a genre from a massive, weakly labeled archive?

Case study: the adventure novel in **10,998** digitized French novels (Gallica)

The problem

How do we retrieve a genre from a massive, weakly labeled archive?

Case study: the adventure novel in **10,998** digitized French novels (Gallica)

Why it's hard

- Genre metadata are **sparse and unreliable**
- Boundaries are **porous** (historical, exotic, maritime, colonial...)
- Exhaustive manual annotation is **impossible** at this scale

What This Talk Is About

The problem

How do we retrieve a genre from a massive, weakly labeled archive?

Case study: the adventure novel in **10,998** digitized French novels (Gallica)

Why it's hard

- Genre metadata are **sparse and unreliable**
- Boundaries are **porous** (historical, exotic, maritime, colonial...)
- Exhaustive manual annotation is **impossible** at this scale

The approach

A **two-stage pipeline**:

1. Supervised classifier on a historical seed list
2. Graph-based refinement over the whole archive

The canon

- A tiny fraction of literary production
- Stabilized by school
- What researchers usually study

The archive

- The mass of printed production
- Popular, serial, forgotten fiction
- Weakly described by metadata

How exactly can we retrieve this *Hors d'usage* literature ? (Cohen, 1999)

The Adventure Novel: An Archive-Genre Par Excellence

Periodization (Letourneux, 2010)

- Peak: **1870–1930**
- Industrial and serial genre
- Verne, Feval, Ponson du Terrail, Aimard...

Bibliographic profile

- Massively produced and read
- Largely forgotten today
- **Porous** boundaries: historical, exotic, maritime, colonial...

Periodization (Letourneux, 2010)

- Peak: **1870–1930**
- Industrial and serial genre
- Verne, Feval, Ponson du Terrail, Aimard...

Bibliographic profile

- Massively produced and read
- Largely forgotten today
- **Porous** boundaries: historical, exotic, maritime, colonial...

The methodological challenge

How can we **retrieve** thousands of adventure novels from a digital archive where:

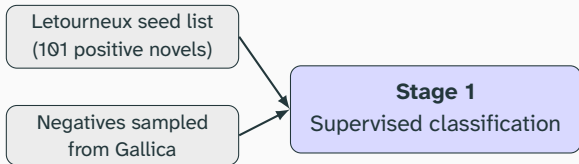
- Genre metadata is **sparse**
- The genre is **bibliographically unstable**
- Exhaustive annotation is **impossible?**

Two-Stage Pipeline

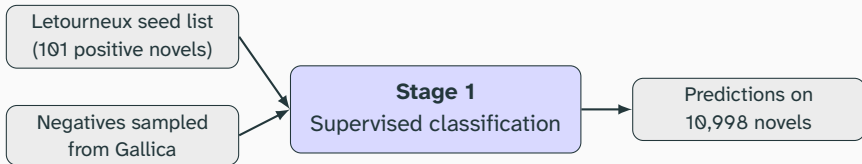
Letourneux seed list
(101 positive novels)

Negatives sampled
from Gallica

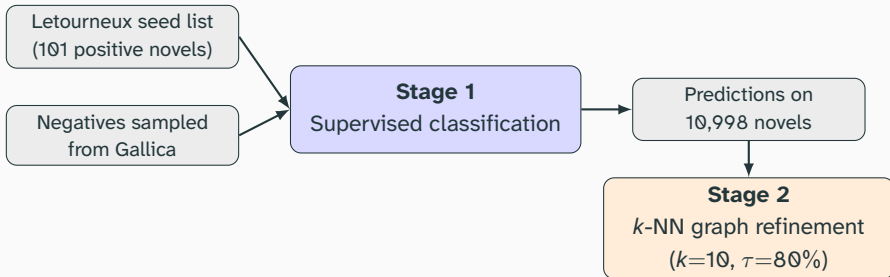
Two-Stage Pipeline



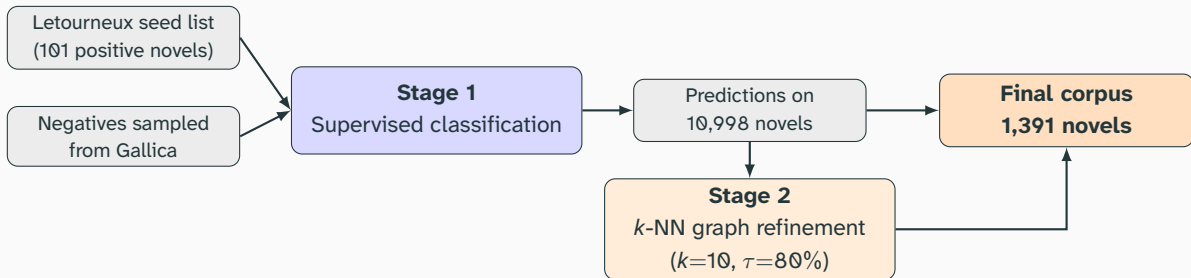
Two-Stage Pipeline



Two-Stage Pipeline



Two-Stage Pipeline



Positives: the Letourneux seed list

- **101** adventure novels cited by Letourneux (2010)
- **36 distinct authors**, max 5 novels per author
- Diversified: canonical names (Verne, Dumas, Feval) *and* forgotten ones (Boussenard, Ferry, Assolant...)

Stage 1: A *Historically Situated* Training Set

Positives: the Letourneux seed list

- **101** adventure novels cited by Letourneux (2010)
- **36 distinct authors**, max 5 novels per author
- Diversified: canonical names (Verne, Dumas, Feval) *and* forgotten ones (Boussenard, Ferry, Assolant...)

Negatives

- Non-adventure novels sampled from Gallica
- **Spread across the full temporal range**, including 1650–1870 and 1930–1950
- Total: **205 training novels**

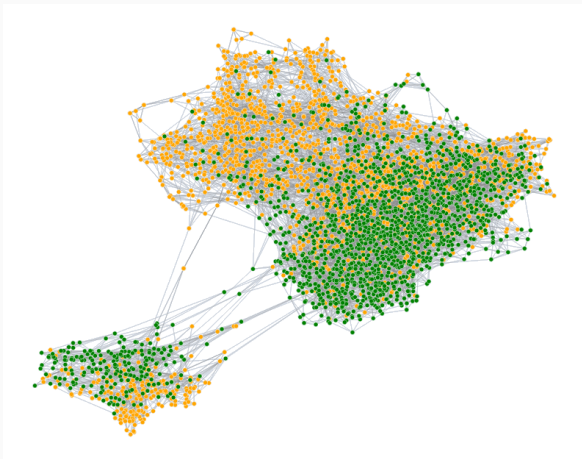
Guardrail: a model that only learned period would fail on these out-of-period negatives.

5-Fold Cross-Validation on 205 Novels

Classifier	Representation	BAcc	F1	AUC
Linear SVM	BoW	.87 $\pm$.04	.87 $\pm$.04	.94 $\pm$.03
Random Forest	BoW	.88 $\pm$.06	.88 $\pm$.06	.95 $\pm$.03
HistGradBoost	BoW	.85 $\pm$.05	.85 $\pm$.05	.94 $\pm$.02
Linear SVM	Emb	.86 $\pm$.05	.85 $\pm$.06	.93 $\pm$.02
Random Forest	Emb	.89 $\pm$.05	.89 $\pm$.05	.95 $\pm$.03
HistGradBoost	Emb	.88 $\pm$.04	.88 $\pm$.04	.95 $\pm$.03

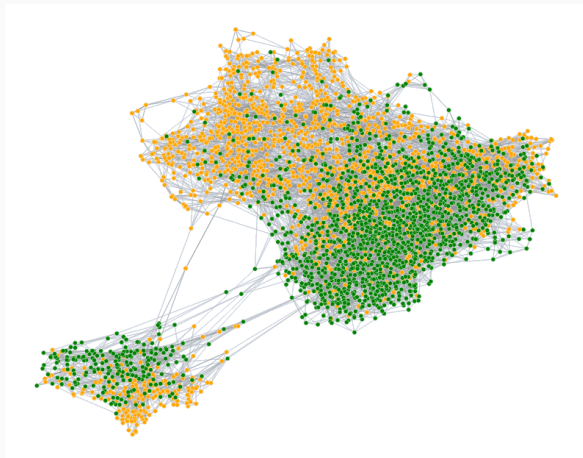
- **All models** exceed .85 BAcc and .93 AUC
- But internal CV is **not sufficient**: the seed is small and historically situated

Stage 2: Neighborhood-Based Refinement



k -NN graph: predicted-adventure candidates (orange)
vs. randomly sampled non-adventure novels (green)

Stage 2: Neighborhood-Based Refinement

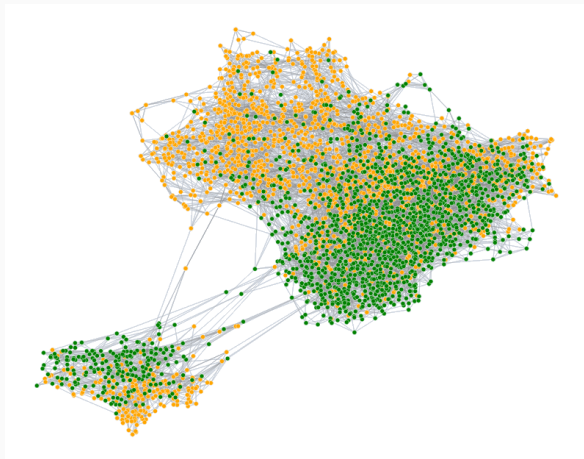


k -NN graph: predicted-adventure candidates (orange)
vs. randomly sampled non-adventure novels (green)

The intuition

If family resemblance holds, an adventure novel should have a **higher-than-chance proportion** of co-labeled neighbors.

Stage 2: Neighborhood-Based Refinement



k -NN graph: predicted-adventure candidates (orange)
vs. randomly sampled non-adventure novels (green)

The intuition

If family resemblance holds, an adventure novel should have a **higher-than-chance proportion** of co-labeled neighbors.

Working postulate

Family resemblance $\approx$ **local label agreement** in the embedding space.

Construction

Directed k -NN graph over the 10,998 novels, cosine similarity on mean paragraph embeddings.

Construction

Directed k -NN graph over the 10,998 novels, cosine similarity on mean paragraph embeddings.

Outlier removal

A novel **predicted as adventure** is **removed** if $\geq 80\%$ of its k neighbors are predicted non-adventure.

$\Rightarrow$ Candidate *false positives*

The Correction Rule

Construction

Directed k -NN graph over the 10,998 novels, cosine similarity on mean paragraph embeddings.

Outlier removal

A novel **predicted as adventure** is **removed** if $\geq 80\%$ of its k neighbors are predicted non-adventure.

$\Rightarrow$ Candidate *false positives*

Infiltrator addition

A **non-predicted** novel is **added** if $\geq 80\%$ of its k neighbors are predicted adventure.

$\Rightarrow$ Candidate *false negatives*

The Correction Rule

Construction

Directed k -NN graph over the 10,998 novels, cosine similarity on mean paragraph embeddings.

Outlier removal

A novel **predicted as adventure** is **removed** if $\geq 80\%$ of its k neighbors are predicted non-adventure.

$\Rightarrow$ Candidate *false positives*

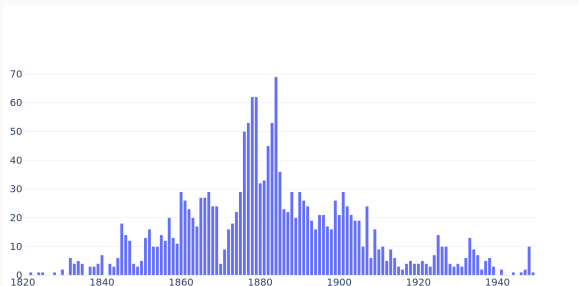
Infiltrator addition

A **non-predicted** novel is **added** if $\geq 80\%$ of its k neighbors are predicted adventure.

$\Rightarrow$ Candidate *false negatives*

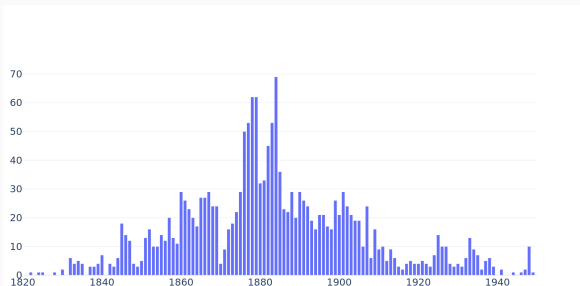
Parameters: $k = 10$, threshold $\tau = 80\%$ – deliberately simple and interpretable.

On Gallica: Temporal Distribution



Temporal distribution of adventure predictions

On Gallica: Temporal Distribution



Temporal distribution of adventure predictions

Convergence with Letourneux

- Peak: **1870–1930** – Matches Letourneux's *independent* periodization

Graph correction

1,375 → **1,391**
adventure candidates

Reclassifications

- **113** outliers removed
- **129** infiltrators added
- **242** total changes

—113 outliers removed

- **Non-fiction:** guidebooks, popular science
- **Parodies:** Allais, *Le Captain Cap*; *Baron Munchausen*
- **Adjacent genres:** Roussel, Gautier, Staël

Qualitative Cases: What the Graph Corrects

–113 outliers removed

- **Non-fiction:** guidebooks, popular science
- **Parodies:** Allais, *Le Captain Cap*; *Baron Munchausen*
- **Adjacent genres:** Roussel, Gautier, Staël

+129 infiltrators added

- **Verne recovered:** *Michel Strogoff*, *Le Chancellor...*
- **Missed subgenres:** Mayne Reid (frontier), Marryat (maritime), Aimard (colonial)
- **Style** diverges from the prototype – *neighborhoods* don't

Qualitative Cases: What the Graph Corrects

–113 outliers removed

- **Non-fiction:** guidebooks, popular science
- **Parodies:** Allais, *Le Captain Cap*; *Baron Munchausen*
- **Adjacent genres:** Roussel, Gautier, Staël

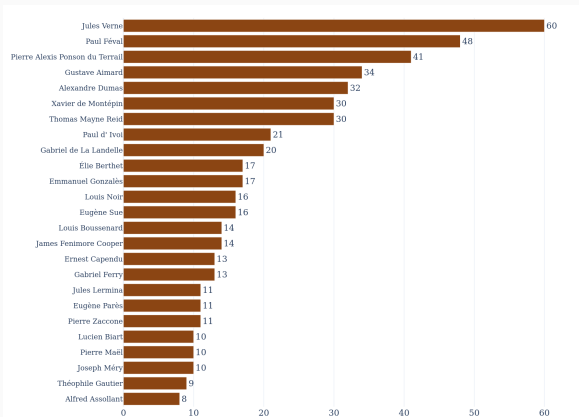
+129 infiltrators added

- **Verne recovered:** *Michel Strogoff*, *Le Chancellor...*
- **Missed subgenres:** Mayne Reid (frontier), Marryat (maritime), Aimard (colonial)
- **Style** diverges from the prototype – *neighborhoods* don't

What this reveals

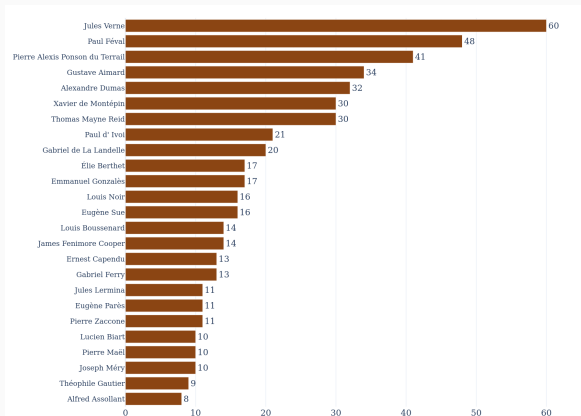
The model captures the **material** of the genre (vocabulary, topoi)
but not its **modalization** (stance, irony, parody)

The Final Corpus: Canonical Authors and the Long Tail



Top 25 authors in the refined adventure corpus

The Final Corpus: Canonical Authors and the Long Tail

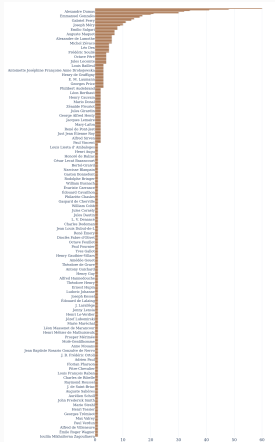


Top 25 authors in the refined adventure corpus

Headliners

- Verne: 60 titles
- Feval: 48 titles
- Ponson du Terrail: 41 titles

The Final Corpus: Canonical Authors and the Long Tail



Titles per author – all 569 authors

Headliners

- Verne: 60 titles
- Feval: 48 titles
- Ponson du Terrail: 41 titles

But above all: the long tail

- **68.5%** of authors: a single title
- Together: **27.6%** of the corpus

The genre is not reducible to a few major names – it is also a **mass production** largely lost to memory.

→ **Genre tags can be retrieved automatically**

A key notion in literary studies, recovered at archive scale

→ **Two stages, supervised + unsupervised**

Refining labels through their relation to the corpus

→ **The k -NN graph is not a stronger classifier**

An archive-refinement heuristic, not a universal method

→ **Corpus construction is interpretive**

Added/removed cases expose the boundary zones (parody, divergent subgenres)

Thank you!

Questions?

Contact

jean.barre@ens.psl.eu

jeanbarre.eu

Corpus, code, full lists:

github.com/crazyjeannot/In-Search-for-Lost-Adventure-Novels

New release

Hors d'usage: Curated French Fiction from Gallica, 1800–1950

Zenodo

Texts, metadata, documentation,
and corpus-construction notes

