A Growing Literature of the Public Sphere: Fiction in Danish Newspapers (1666-1850)

Pascale Feldkamp*, Alie Lassche*, Rie Eriksen*, Kit Morgenstjerne*, Johan Heinsen‡, Kristoffer Nielbo*, Yuri Bizzoni*

*Center for Humanities Computing, Aarhus University, ‡MASSHINE, Aalborg University

{pascale.feldkamp, a.w.lassche, yuri.bizzoni}@cas.au.dk, heinsen@dps.aau.dk

Abstract

Digitized literary corpora of the 19th century largely focus on standalone volumes, sidelining the broader and more diverse literary production of the period. Fiction published in less enduring formats – such as novellas and serialized pieces in newspapers – remains underexplored, particularly for low-resource languages like Danish, despite the growing availability of digitized newspaper archives. This paper addresses that gap by identifying and tagging fiction in Danish newspapers (1666–1850). We (1) present a manually annotated dataset of 1,831 articles with both binary (fiction/nonfiction) and fine-grained subcategories (travelogue, biography, essay), and (2) evaluate a document-embedding classifier that achieves an F1-score of up to 0.89 for the fiction/nonfiction distinction. Building on this pipeline, we further provide two resources for future research: (a) fiction probability scores for nearly five million newspaper articles ($n = 4,898,084$), and (b) a small, cleaned, and curated subset of newspaper fiction ($n = 139$), intended as a growing resource.

Keywords: literature, historical newspapers, embeddings, serialized fiction

1. Introduction and Related Works

Only a small fraction of novels are widely studied, while most of literary production remains what Franco Moretti (2000) famously called "the great unread". Recent computational and statistical approaches have begun to expand the literary horizons (Underwood, 2019; Algee-Hewitt et al., 2016), examining lesser-known texts and opening the way for a kind of literary sociology that traces the circulation and diversity of literary production.

However, this ambition falls short in practice. In most digitized corpora – especially for under-resourced languages like Danish – the standalone novel remains the dominant form, overshadowing the diverse genres and formats that animated the 19th-century print market (Stangerup, 1936).[1] As the literary market consolidated (Bourdieu, 1996), much literary culture circulated through ephemeral media: newspapers, periodicals, and serials. The 19th century marked a turning point: literature became a mass-cultural product in the everyday media landscape (Easley, 2024; Horstbøll, 1999). While newspapers were central to the formation of the public sphere and the emerging nation-state (Habermas, 1989; Anderson, 2006), they also provided a venue for literary experimentation. Serialized fiction circulated across and beyond national borders, connecting readers, creating new genres, and attracting subscribers (Lehrmann, 2018). But despite some recent digital initiatives[2], there are few available datasets of these more impermanent forms of 19th-century literature. This is a critical gap. When writing literary history, we lose a great deal of information about the real scope of literary production, but we also lose an opportunity to understand literature's role in public life during this formative period of the emerging public sphere and nation-state. If most corpora overlook the transient life of literature, newspapers offer a way to restore it. Luckily, recent years have seen a significant digitization effort for historical newspapers. Detecting fiction directly within them allows us to glimpse the everyday circulation of stories that once animated public life and to recover lost dimensions of 19th-century literary culture.

Recent efforts have effectively distinguished fiction and nonfiction (Qureshi et al., 2019), also in the noisy environments of historical newspapers (Feldkamp et al., 2025; Repo, 2024). That is despite the task being theoretically complex: the distinction between fiction and nonfiction is notoriously difficult to pin down, with some arguing it depends more on reader framing than textual profile (Culler, 2002; Fish, 2003; Stockwell, 2002). This ambiguity is especially pronounced in historical sources: 19th-century literature and journalism competed to depict social reality and assert social truths (Lepenies

[1] Many corpora index novels published as standalone volumes, such as the Chicago Corpus, the ELTEC corpora, or Common Library 1.0; and, for Danish, the MeMo corpus (Bjerring-Hansen et al., 2022). In contrast, computational studies of *modern* (and English) literature focus on exploring alternative, impermanent forms such as Wattpad and fanfiction (Pianzola et al., 2020; Jacobsen and Kristensen-McLachlan, 2025).

[2] For instance, the Ciphers project: https://libraryponders.github.io/index.html

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 40–49

and Plard, 1995), while the modern ideal of journalistic objectivity emerged gradually (Schudson, 2001). Writers like Zola moved between literary and journalistic modes, and narrative techniques circulated across registers, with 'realism' shaping early novelistic forms (Watt, 2001). Furthermore, authorization strategies – claims of eyewitness accounts, discoveries of hidden documents, and use of documentary conventions – have long been used to enhance verisimilitude in fiction (Panayotakis et al., 2010). Nevertheless, linguistic research identifies features distinguishing fiction from nonfiction, including adjective/adverb ratios, pronouns, type-token ratios, nominalizations, and syntactic complexity, with nonfiction generally exhibiting higher information density (Qureshi et al., 2019; Kazmi et al., 2022; Kubát and Milička, 2013; Sadeghi and Dilmaghani, 2013; Vicente et al., 2021; Dijk, 2009). Recent studies show that semantic document embeddings outperform surface and sentiment features in detecting narrative segments (Repo, 2024; Laippala et al., 2019; Feldkamp et al., 2025), suggesting that fiction's profile is not only stylistic or sentiment-based but also semantic.

Here, we present a pipeline for extracting fiction from noisy historical Danish newspapers using text embeddings, along with a small, curated set of serialized fiction as a proof of concept.[3]

2. Methods

2.1. Data

2.1.1. Newspapers

The corpus used in this study consists of Danish newspapers published in the conglomerate state of Denmark-Norway between 1666 and 1850.

Due to OCR difficulties with *fraktur*, thin paper, and complex layouts, the corpus was re-digitized by the ENO group at Aalborg University[4] using custom transcription models in Transkribus (Kahle et al., 2017). Articles were segmented via line-level classification using `SetFit` and `RandomForest` models. The resulting corpus spans 28 Danish newspapers, printed in both Danish and Norwegian towns, covering both center and periphery, with a total of almost five million individual articles.

Notably, the corpus includes early examples of children's literature in the *Adresseavis for Børn* (1779-1782).[5] This was the first Danish newspaper explicitly aimed at a young readership, featuring a variety of genres including moral tales, travelogues, didactic essays, letters, and brief news items. While most of the newspaper's content consists of short pieces, a smaller part follows a serial model, with stories spanning several editions. We hypothesize that the inclusion of *Adresseavis for Børn* broadens the stylistic and social range of the dataset, as it may exhibit a distinct topical and linguistic profile compared to the rest of the corpus.

2.1.2. Annotated set

The articles for annotation were partly randomly selected from the entire period, and partly gathered to capture serialized novels, with batches of fiction and nonfiction identified using search terms such as "to be continued".[6] Each article was tagged by at least two expert annotators.[7] In the annotated set, we tagged both the coarse fiction/nonfiction distinction and fine-grained categories ('biography', 'travelogue', 'essay', 'poem', 'anecdote', 'narrative nonfiction', 'play'). To focus on the main distinction, we excluded hybrid or distinct forms such as 'narrative nonfiction', 'anecdote', 'poetry', and 'play'. Still, to maintain task complexity, we retained subcategories 'biography' and 'travelogue' that occur under both fiction and nonfiction, as annotators could reliably distinguish the higher-level register.[8] Only articles with agreement from at least two annotators were kept, resulting in 1,962 tagged articles, of which 1,831 exceed 20 words. They span 1759–1874.[9]

When clear story signals – titles, tags, prologues, formatting – were present in the inspected newspaper scans, texts were annotated as fiction; otherwise, annotation focused on distinguishing registers. Given the fluid boundary between early news and fiction, fiction is arguably better treated as a register rather than a fixed category (Repo, 2024), consistent with literary theory that frames literariness as a mode of communication rather than a bounded product (Jakobson, 1981; Bachtin et al., 2014; Jauss and Benzinger, 1970). So in cases where there was no clear signal but where, for ex-

[3] The anonymized repository for the code underlying this paper is available here: `https://github.com/centre-for-humanities-computing/feuilleton_dataset`, including links to all resources presented in this paper.

[4] `https://hislab.quarto.pub/eno/`.

[5] Later called *Avis for Børn* (Newspaper for Children).

[6] Since the pipeline aims to identify literary segments as they appear in practice, we retained editorial markers like "to be continued", as they likely reflect conditions of downstream application.

[7] We had three annotators, two with a university background in Literary Studies and one in the Study of Religion. The annotators were trained in close reading and had comprehensive knowledge of textual culture in 18th and 19th-century Denmark.

[8] The annotated set is available here: `https://huggingface.co/datasets/chcaa/fiction-nonfiction-testset`

[9] This period reflects peak newspaper output, not annotation choices.

Features	Class	Precision	Recall	F1-score
TF-IDF (max. 5,000)	*Fiction*	0.89 ± 0.03	0.85 ± 0.06	0.87 ± 0.02
	Nonfiction	0.87 ± 0.04	0.91 ± 0.04	0.88 ± 0.01
Document embeddings (768 dimensions)	*Fiction*	0.88 ± 0.02	0.88 ± 0.05	0.88 ± 0.03
	Nonfiction	0.89 ± 0.04	0.89 ± 0.02	0.89 ± 0.02

Table 1: Average classification performance and SD across 5 folds. Precision, recall, and F1-score are reported per class.

Category	N articles	N words
All	1,831	569,543
Nonfiction	951	225,146
Fiction	880	344,397
Biography	170	66,017
Travelogue	80	28,786
Essay	78	32,285

Table 2: Description of the annotated set (1759-1874). Note that these are the numbers after we removed articles with less than 20 words. A lot of fiction articles have the general fiction tag.

ample, a biographical text leaned heavily on literary devices (such as first-person or internal focalization, storylines, tropes, etc.) that were characteristic for fiction in the period, we annotated it as fiction. This is complicated by article fragmentation, which was overcome by filtering out short texts and, finally, on annotator agreement.

2.2. Classification

Embeddings: We tested six embedding models for the task of differentiating fiction and nonfiction against a baseline of TF-IDF representations (max. 5,000 features), summarized in Appendix subsection 3.5. The best performing model was `Old_News_Segmentation_SBERT_V0.1`,[10] which is why we used it for making document embeddings.[11]

Model: Some fiction texts (longer running pieces) appeared as multiple article fragments or feuilletons serialized across editions. We assigned a unique ID to each serialized piece or fragment and used dummy values for missing or incomplete IDs to maintain consistent grouping. For evaluation, we applied a `StratifiedGroupKFold` cross-validation scheme that preserved both class balance and ID integrity – ensuring that fragments or installments of the same piece were never split between train and test sets. Using 5-fold cross-validation, we trained a `LogisticRegression` classifier with balanced class-weight to account for label imbalance.[12] The classifier was trained on each fold and evaluated on the held-out set, computing precision, recall, and F1-scores per class and averaging across folds. This procedure was applied to both TF-IDF features and embeddings. Results are shown in Table 1. Note how closely embeddings and TF-IDF perform, suggesting that both semantic and lexical patterns are informative for the fiction/nonfiction distinction.[13]

3. Results

3.1. Tagged corpus

We applied the best-performing (embedding-based) model to all articles in the newspaper corpus (1677-1849) and assigned fiction probability scores. To get a sense of the distribution of fiction tags across the corpus and how the makeup of the newspaper landscape (high/low heterogeneity) changes it, we plotted percentages of fiction tags across the period (Figure 1).

Importantly, among articles tagged as fiction ($n = 77,513$; probability > 0.5), predicted probabilities show no correlation with publication date or text length.[14] While fiction tags appear more common in earlier newspapers (see Figure 1), this is unlikely to reflect model bias, as the training set included no examples before 1750. The spike in the 1680s is better explained by the newspaper *Danske Mercurius* (the first red spike in Figure 1), which presented news in flowing alexandrines accompanied by short poetic reflections. Upon manual inspection, it seems our classifier assigns high fiction probability to poetry, but also tended to misclassify obituaries as fiction, perhaps for their distinct poetic and emotional tone. The misclassifications

[10] https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0. It was developed on the same Danish historical newspaper corpus for article segmentation.

[11] Procedure as outlined for the model comparison in Appendix subsection 3.5: texts were split into max input chunks, and a mean embedding was computed by averaging chunk embeddings.

[12] Stratification and classifier from Scikit-learn.

[13] Downscaling TF-IDF to 768 features or selecting the top 768 using `SelectKBest` with a chi-squared (χ^2) test yields similar results, indicating that TF-IDF performance is not simply due to its larger feature space.

[14] Spearman's $\rho < 0.07$, $p > 0.5$.

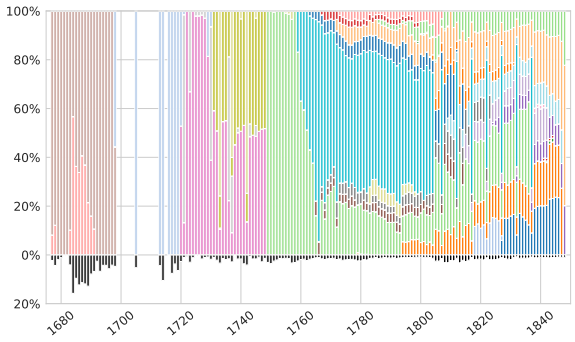

Figure 1: Distribution of newspapers as a percentage of total articles per year (top, in color) and histogram of the share of articles labeled fiction per year (bottom). Fiction is slightly more prominent in earlier newspapers, but overall remains low and stable (average 1.47%), despite a more heterogeneous publishing landscape.

highlight both a limitation and a strength of our approach to annotating fiction as a matter of literary register. While the model foregoes publishing categories, it picks up on linguistic mode – stylistic and tonal cues that cross formal category boundaries. Remember that we purposely excluded poetry and hybrid categories like anecdotes and narrative nonfiction. As such, rather than pointing to the model, these results seem to tell us something about formal categories, e.g., that poetry or obituaries appear in a fiction register.

3.2. Curated dataset

To better understand serialized fiction and ensure the reliability of story-level analyses, we created a curated dataset by manually inspecting a subset of high-probability predictions (probability > 0.99). The threshold was chosen to maximize confidence in the classifier's assignments, since fully automated grouping across installments and article fragments remains challenging. From this subset, we collected all fragments of the same story across the corpus and grouped serialized fiction into installments (a, b, c, etc.), assigning each story a unique ID. The resulting dataset contains 139 unique IDs, of which 69 consist of multiple installments (see Table 3). Additional cleaning and tagging ensured stories were spell-checked and assigned to general or more specific categories.[15] Some categories, like installments from the same series or travelogue and biography, show clustering (see Appendix Figure 3). Children's literature, by contrast, is widely dispersed, indicating it does not form a semantically consistent category in our data. While relatively small, this curated set serves as a proof-of-concept resource, allowing us to explore serialized fiction at the story level and providing a foundation for future, larger-scale expansions.

General	
N words	361,344
N articles	266
N stories	139
N stories $>$ 1 part	69
Stories per category	
Biography	35
Travelogue	18
Lovestory	12
Children's literature	17
Dialogue	2
Satire	2

Table 3: Description of the curated set (1763-1841).

3.3. Conclusion and Discussion

Our classification results suggest that fiction and nonfiction in newspapers differ in both word-usage patterns (TF–IDF) and semantic representations. In fact, it is certainly possible to distinguish between the two, even in short texts: probabilities show no correlation with length, suggesting that short literary fragments are just as recognizably "fictional" as longer ones. This makes the approach promising for other newspaper datasets. Although the share of fiction in any given year is modest, it's share is stable across time and still represents a substantial body of material ($\sim$77,500 articles).

The curated dataset illustrates both the feasibility of extracting literary corpora from historical newspapers and the character of the literary culture they contain, shaped by translation, transmission, and republication. Many stories are translations – from German, English, and French – and are frequently republished across newspapers, pointing to active (transnational) exchange. While Danish novelistic literature may have remained largely insular, newspaper fiction participated in wider literary circuits (Lehrmann, 2018). Authorship and transmission were often fluid: names are frequently pseudonymized, abbreviated, or absent, and the provenance of many texts is uncertain. For example, *The Maidens War* derives from a German source claiming Latin origins, while *The Labyrinth* allegedly comes from Zend, though the originality of such texts was questioned even at the time (Rask, 1826). This anonymity and cross-border mobility reflect the ephemeral, fast-moving nature of the medium. Where authorship can be identified, ca. 20% of names are female, matching proportions in contemporaneous novel publication (Degn et al., 2025), though translators – often women – were frequently uncredited (Nøding, 2017).

[15] This dataset is available here: https://huggingface.co/datasets/chcaa/Press-and-Plot

Unlike long, standalone volumes, these stories seem designed to move across authors, languages, and borders as smoothly as possible. Transmission and republication strategies reveal a highly dynamic, transnational literary space that both connected and shaped the emerging public sphere. Future research could trace these trajectories to better understand how fiction contributed to cultural exchange, reading practices, and the formation of modern national literary life.

Ethical considerations & limitations

Several limitations should be noted. First, annotator subjectivity remains a factor: even when depending on agreement and textual signals in the original scans, the fiction/nonfiction distinction is partly interpretive, reflecting the fluidity of literary registers in historical newspapers. Second, temporal coverage is restricted to 1666–1850; textual profiles may differ before and after this period. Notably, fiction reportedly increased toward the late 19th century, coinciding with the expansion of printed mass communication and the literary market (Feldkamp et al., 2024; Horstbøll, 1999). Future extraction of fiction in late-19th-century newspapers could therefore yield particularly rich literary corpora.

Historical newspapers reflect the socio-political biases of their time, including gender, class, and colonial perspectives. At the textual level, language referring to marginalized groups does not meet modern standards and should be understood in its historical context. At the production level, many female translators remain uncredited (Nøding, 2017), reflecting persistent inequities in attribution. While the dataset is historical and publicly available – minimizing privacy/copyright concerns – researchers should engage with it carefully, remaining attentive to both historical context and representational biases.

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. *Canon/Archive. Large-scale Dynamics in the Literary Field*. Stanford Literary Lab.

Benedict Anderson. 2006. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.

Michail Michajlovič Bachtin, Michael Holquist, and Vern McGee. 2014. *Speech Genres and Other Late Essays*. University of Texas Press, Austin.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending Fractured Texts. A heuristic procedure for correcting OCR data: 6th Digital Humanities in the Nordic and Baltic Countries Conference, DHNB 2022. In *CEUR Workshop Proceedings*, volume 3232, pages 177–186, Uppsala, Sweden.

Pierre Bourdieu. 1996. *The Rules of Art: Genesis and Structure of the Literary Field*. Stanford University Press.

Jonathan D. Culler. 2002. Literary competence. In *Structuralist poetics: structuralism, linguistics and the study of literature*, pages 131–152. Routledge, London. OCLC: 56560333.

Kirstine Nielsen Degn, Jens Bjerring-Hansen, Ali Al-Laith, and Daniel Hershcovich. 2025. Unhappy Texts?: A Gendered and Computational Rereading of the Modern Breakthrough. *Scandinavian Studies*, 97(1):1–24.

Teun A. van Dijk. 2009. *News as discourse*. Routledge, New York. OCLC: 868975895.

Alexis Easley, editor. 2024. *British writers, popular literature and new media innovation, 1820-45*. Nineteenth-century and neo-Victorian cultures. Edinburgh University Press, Edinburgh.

Pascale Feldkamp, Alie Lassche, Katrine Frøkjær Baunvig, Kristoffer Nielbo, and Yuri Bizzoni. 2025. Fact from fiction: Finding serialized novels in newspapers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 695–707, Vienna, Austria. Association for Computational Linguistics.

Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig, and Kristoffer Nielbo. 2024. Canonical status and literary influence: A comparative study of Danish novels from the modern breakthrough (1870–1900). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 140–155, Miami, USA. Association for Computational Linguistics.

Stanley Eugene Fish. 2003. *Is there a text in this class? the authority of interpretive communities*, 12. print edition. Harvard Univ. Press, Cambridge, Mass.

Jürgen Habermas. 1989. *The structural transformation of the public sphere : an inquiry into a category of bourgeois society*. MIT Press.

Henrik Horstbøll. 1999. *Menigmands medie: det folkelige bogtryk i Danmark 1500 - 1840: en kulturhistorisk undersøgelse*. Number 19 in Danish Humanist Texts and studies. Det Kongelige

Bibliotek, Museum Tusculanums Forlag, København.

Mia Jacobsen and Ross Deans Kristensen-McLachlan. 2025. Beyond Style: Rethinking Computational Fanfiction Research. *Journal of Data Mining & Digital Humanities*, NLP4DH:16414.

Roman Jakobson. 1981. Linguistics and poetics. In *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton.

Hans Robert Jauss and Elizabeth Benzinger. 1970. Literary History as a Challenge to Literary Theory. *New Literary History*, 2(1):7.

P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Arman Kazmi, Sidharth Ranjan, Arpit Sharma, and Rajakrishnan Rajkumar. 2022. Linguistically Motivated Features for Classifying Shorter Text into Fiction and Non-Fiction Genre. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 922–937, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Miroslav Kubát and Jiří Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, 20(4):339–349.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297, Turku, Finland. Linköping University Electronic Press.

Ulrik Lehrmann. 2018. Føljetonromanen og dansk mysterie-litteratur i 1800-tallet. *Passage - Tidsskrift for litteratur og kritik*, 33(79):31–46. Number: 79.

Wolf Lepenies and Henri Plard. 1995. *Les trois cultures - entre science et littérature, l'avènement de la sociologie*, 0 edition edition. MSH PARIS, Paris.

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

Aina Nøding. 2017. *Periodical Fiction in Denmark and Norway before 1900*. Oxford University Press.

Stelios Panayotakis, Maaike Zimmerman, and Wytse Hette Keulen, editors. 2010. *The ancient novel and beyond*. Number 241 in Mnemosyne, bibliotheca classica Batava 0169-8958. Supplementum. Brill, Leiden, Netherlands Boston.

Federico Pianzola, Simone Rebora, and Gerhard Lauer. 2020. Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS ONE*, 15(1):e0226708. Publisher: Public Library of Science.

Mohammed Rameez Qureshi, Sidharth Ranjan, Rajakrishnan Rajkumar, and Kushal Shah. 2019. A simple approach to classify fictional and non-fictional genres. In *Proceedings of the Second Workshop on Storytelling*, pages 81–89, Florence, Italy. Association for Computational Linguistics.

Rasmus Rask. 1826. *Om Zendsprogets og Zendavestas Ælde og Ægthed*. Andreas Seidelin.

Liina Repo. 2024. Towards automatic register classification in unrestricted databases of historical English. In *Linguistics across Disciplinary Borders: The March of Data*, 1 edition, pages 97–126. Bloomsbury Publishing Plc.

Karim Sadeghi and Sholeh Karvani Dilmaghani. 2013. The Relationship between Lexical Diversity and Genre in Iranian EFL Learners' Writings. *Journal of Language Teaching and Research*, 4(2):328–334.

Michael Schudson. 2001. The objectivity norm in American journalism. *Journalism*, 2(2):149–170. Publisher: SAGE Publications.

Hakon Stangerup. 1936. *Romanen i Danmark: Romanen i det Attende Århundrede*. Levin & Munksgaards Forlag.

Peter Stockwell. 2002. *Cognitive poetics: an introduction*. Routledge, London.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago, IL.

Marta Vicente, María Miró Maestre, Elena Lloret, and Armando Suárez Cueto. 2021. Leveraging Machine Learning to Explain the Nature of Written Genres. *IEEE Access*, 9:24705–24726.

Ian Watt. 2001. *Rise of the Novel, Updated Edition*. University of California Press, Berkeley, CA.

Appendix A

3.4. Distribution of articles over time

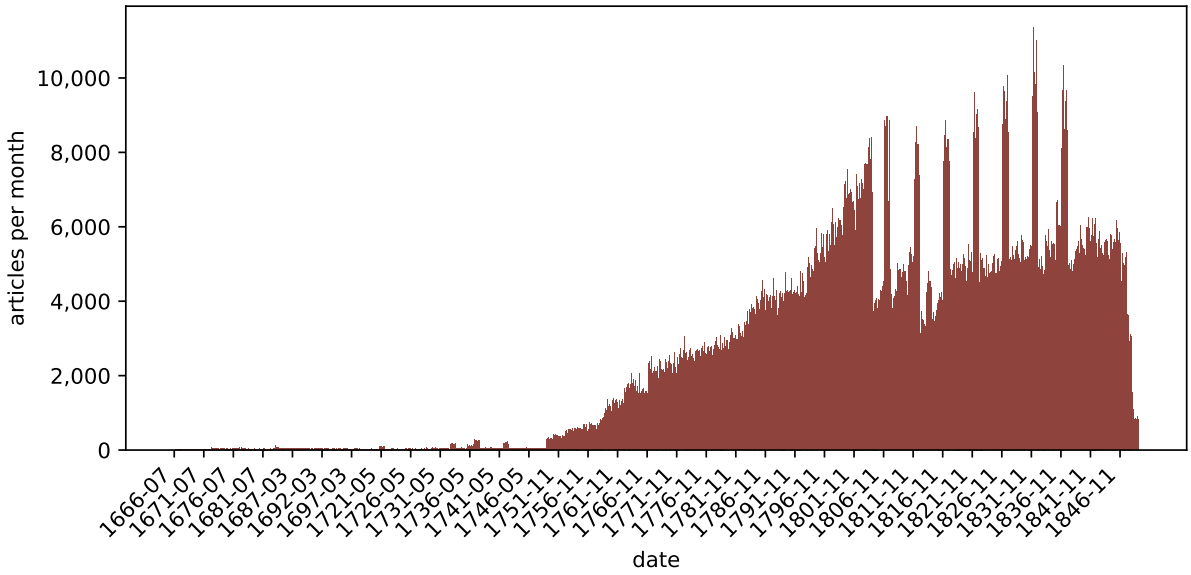

Figure 2: Distribution of the full corpus, number of articles per month.

3.5. Model comparison

Model	Precision		Recall		F1-score		Accuracy
	fiction	non-fiction	fiction	non-fiction	fiction	non-fiction	
TF/IDF	0.887	0.850	0.840	0.905	0.860	0.874	0.869
`MeMo-BERT-03`	0.873	0.878	0.878	0.875	0.875	0.876	0.876
`Old_News`	0.884	0.886	0.886	0.887	0.885	0.886	0.885
`e5`	0.888	0.876	0.872	0.892	0.879	0.884	0.882
`jina`	0.863	0.868	0.868	0.865	0.865	0.867	0.866
`bge-m3`	0.861	0.871	0.869	0.864	0.864	0.867	0.866
`gemma`	0.816	0.813	0.808	0.821	0.812	0.817	0.816

Table 4: Performance (averaged across 5 folds) of TF/IDF and six embedding models on the fiction/non-fiction classification task, evaluated using 5-fold `StratifiedGroupKFold` cross-validation with group-preserving splits. Metrics reported are precision, recall, F1-score (per class), and overall accuracy. Note that the second-best model (`e5`) sometimes has a higher precision or recall for one of the classes, while `Old_News` performs more consistently (i.e., has slightly higher F1 for fiction).

Of the models tested, three had shown potential in earlier studies with Danish historical corpora: among these, two were fine-tuned on nineteenth-century Danish texts, and one was multilingual. Full model names and URLs are shown in Table 5. We selected three other state-of-the-art multilingual models for their strong performance in the Multilingual Text Embedding Benchmark (MTEB), manageable size ($< 1B$ parameters), non-instruction-tuned nature, and high maximum input length (8,194 tokens).

Pooling: For all models except `jina-embeddings-v3`, `bge-m3` and `embeddinggemma-300m`, the maximum input length was limited to 512-514 tokens. In all cases, each feuilleton text was split into chunks of up to the maximum number of tokens, and a mean embedding was computed by averaging the resulting chunk embeddings.

Model	Max tokens	Dimensions	Layers	Source
MeMo-BERT-03	514	768	12	https://huggingface.co/MiMe-MeMo/MeMo-BERT-03
Old_News_Segmentation_SBERT_V0.1	512	768	12	https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0.1
bge-m3	8,192	1,024	24	https://huggingface.co/BAAI/bge-m3
embeddinggemma-300m	2,048	768*	24	https://huggingface.co/google/embeddinggemma-300m
jina-embeddings-v3	8,192	1,024*	24	https://huggingface.co/jinaai/jina-embeddings-v3
multilingual-e5-large	514	1,024	24	https://huggingface.co/intfloat/multilingual-e5-large

Table 5: Overview of full model names. We also show the maximum input context length, final embedding dimension size, number of hidden layers, and HuggingFace urls. The order of models is by language (Danish models on top) and alphabetical.

3.6. Classifier comparison

For our final classification with the Old_News embeddings, we tested various classification models. We find that Logistic regression performs consistently well (as does Logistic Elastic Net), which was why we used it for the final tagging of articles.

Classifier	Class	Precision	Recall	F1-score
LogisticRegression	*Fiction*	0.88 ± 0.03	0.88 ± 0.06	0.88 ± 0.02
	Non-fiction	0.89 ± 0.05	0.89 ± 0.04	0.89 ± 0.01
LogisticElasticNet	*Fiction*	0.88 ± 0.03	0.88 ± 0.05	0.88 ± 0.02
	Non-fiction	0.89 ± 0.04	0.89 ± 0.03	0.89 ± 0.01
LinearSVC	*Fiction*	0.87 ± 0.03	0.85 ± 0.05	0.86 ± 0.02
	Non-fiction	0.87 ± 0.04	0.88 ± 0.03	0.87 ± 0.02
RandomForest	*Fiction*	0.90 ± 0.02	0.85 ± 0.06	0.87 ± 0.02
	Non-fiction	0.87 ± 0.04	0.91 ± 0.03	0.89 ± 0.01

Table 6: Average classification performance and standard deviation across 5 folds. Precision, recall, and F1-score are reported per class for each classifier. Logistic Regression and ElasticNet just slightly outperform the others based on F1-score.

3.7. Misclassifications

We include an analysis of the false positives and negatives in our fiction/non-fiction classification. Note that all subcategories can be labeled both fiction or non-fiction in the gold standard. Analysis of misclassifications (Table 7) shows that certain subcategories consistently sit near the boundary of the fiction register. Fictional biographies and travelogues are often misclassified as nonfiction, suggesting that these genres share stylistic features with nonfiction, whereas essays and general nonfiction are occasionally misclassified as fiction, likely because they adopt literary or narrative elements. Overall, the patterns reveal which subcategories most closely overlap with the linguistic and stylistic cues the model associates with fiction, that is, biographies and travelogues make up the largest share of false negatives, while essays account for the largest share of false positives, indicating that the model tends to read fictionalized life-writing as nonfiction and stylistically marked essays as fiction.

Subcategory	False Negatives			False Positives		
	Count	% of subcategory	% of total FNs	Count	% of subcategory	% of total FPs
Biography	37	21.76	34.91	3	1.76	2.91
Travelogue	21	23.86	19.81	12	13.64	11.65
Essay	1	1.28	0.94	24	30.77	23.30

Table 7: False negatives (fiction predicted as non-fiction) and false positives (non-fiction predicted as fiction) by subcategory. Percentages show both the proportion relative to the subcategory and relative to the total number of false negatives/positives.

3.8. Semantic space of the curated set

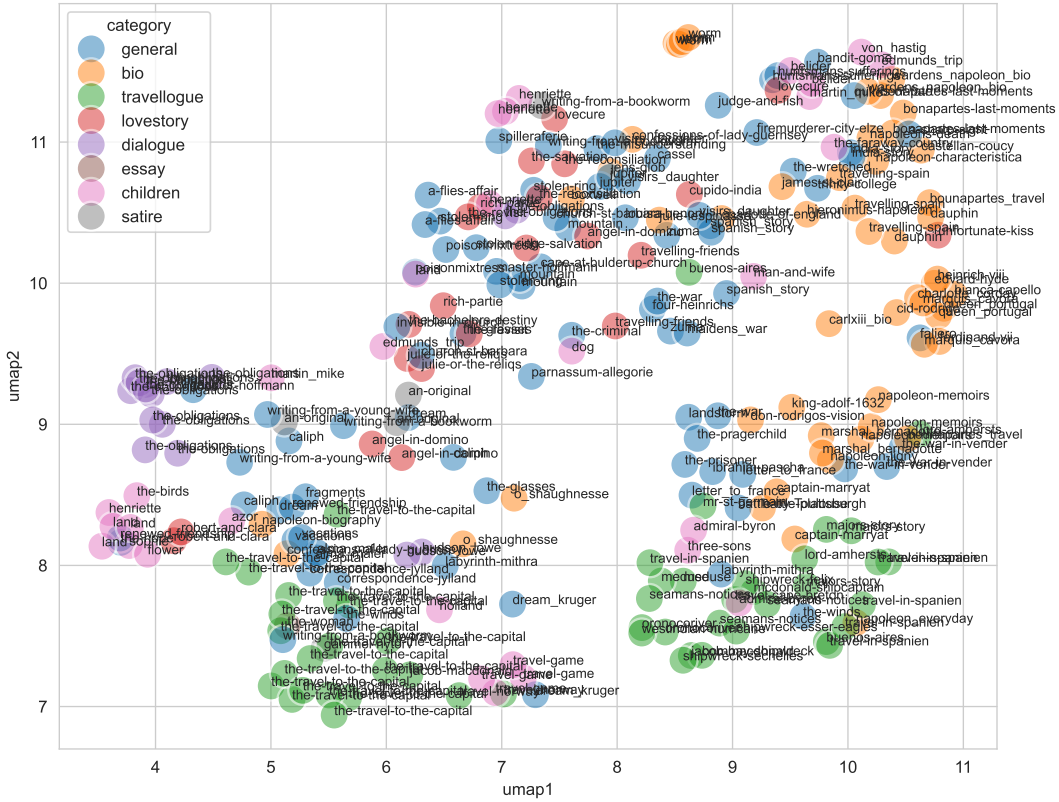

Figure 3: UMAP of the curated set of fiction all individual pieces/installments. Here, we see some tendential clusters. (a) of installments from the same series, and (b) of travelogue and biography tags. For example we see stories involving Napoleon clustering at the upper right corner. However, this is not the rule: consider the relative spread of, e.g., childrens literature across the whole semantic space, indicating that children's literature is not a semantically consistent category (in our data).