Fact from Fiction: Finding Serialized Novels in Newspapers

Anonymous ACL submission

Abstract

Digitized literary corpora of the 19th century favor canonical and novelistic forms, sidelining a broader and more diverse literary production. Serialized fiction – widely read but embedded in newspapers – remains especially underexplored, particularly in low-resource languages like Danish. This paper addresses this gap by developing methods to identify fiction in digitized Danish newspapers (1818–1848). We (1) introduce a manually annotated dataset of 1,394 articles and (2) evaluate classification pipelines using both selected linguistic features and embeddings, achieving F1-scores of up to 0.91. Finally, we (3) analyze feuilleton fiction via interpretable features to test its drift in discourse from neighboring nonfiction. Our results support the construction of alternative literary corpora and contribute to ongoing work on modeling the fiction–nonfiction boundary by operationalizing discourse-level distinctions at scale.[1]

1 Introduction

A significant obstacle for large-scale literary analysis and historiography is that digitized corpora overwhelmingly prioritize familiar genres and canonical works, leaving much of historical literary production underexplored (Algee-Hewitt et al., 2016; Moretti, 2000; Underwood, 2019). This bias is especially pronounced in 19th-century collections, where novels dominate despite a rich ecosystem of genres and publication formats that flourished in the expanding print market (Hertel, 2018; Stangerup, 1936).[2] Among underrepresented but widely read forms are serialized fiction and feuilleton novels – embedded in newspapers rather than published as standalone volumes (Lehrmann, 2018). While traditional scholarship increasingly engages with serialized forms – and some digital efforts have addressed serialization[3] – computational literary studies often focus on accessible, curated, and canonized sources, inadvertently reinforcing existing biases. Digital resources for underrepresented languages like Danish reflect the same tendencies:[4] they often prioritize canonical novels or curated editions of major authors,[5] while alternative forms remain largely inaccessible.

Yet the resources for redressing this imbalance already exist. Danish newspapers from the 19th century have been extensively digitized, offering new opportunities for recovering serialized fiction at scale and (re)writing a more representative, complexity-aware literary history. This material comes with its challenges: digitized newspapers are noisy, with heterogeneous layouts, mixing news items, advertisements, and nonfiction content, with OCR and segmentation errors. Consequently, the first obstacle is methodological: how can we systematically identify fiction in such noisy, heterogeneous environments?

This paper has two goals: first, to test whether classification pipelines based on lexical frequencies, linguistic features, or semantic embeddings can reliably extract fictional from nonfictional discourse in Danish newspapers (1818–1848); and second, to probe language use in feuilleton novels. In both tasks, we contribute to efforts to recover overlooked forms and explore the fiction–nonfiction boundary – a distinction that is theoretically rich but difficult to operationalize (Heyne, 2001; Jakobson, 1981). Our approach helps build literary corpora that better reflect the scale and heterogeneity of 19th-century literary culture.[6]

[1] Our code is available at: `https://anonymous.4open.science/r/factfiction_newspapers-E174`.

[2] Many corpora focus on novels, such as the Chicago Corpus, the ELTEC corpora, or the Common Library 1.0.

[3] Such as the Ciphers project: `https://libraryponders.github.io/index.html`.

[4] E.g., the MeMo corpus: `https://huggingface.co/datasets/chcaa/memo-canonical-novels`.

[5] E.g., Kierkegaard, H.C. Andersen, and Grundtvig.

[6] This research forms part of a Ph.D. project on literary clio-

2 Related works

The boundary between fiction and nonfiction is neither fixed nor purely textual. It is shaped by genre conventions, reader framing (Culler, 2002; Fish, 2003), and historical norms (Heyne, 2001; Schudson, 2001). In the 19th century, this boundary was especially unstable: literature and journalism competed for authority to depict social reality, and hybrid forms like the feuilleton blurred reportage and fiction to assert social truths (Lepenies and Plard, 1995). Writers like Zola moved between literary and journalistic modes, while narrative techniques were widely used in news discourse. The modern journalistic "objectivity" ideal only stabilized gradually over the century (Schudson, 2001).

While today's newspapers more clearly signal truth-claims, many argue a fiction/nonfiction distinction still hinges more on reception than form (Stockwell, 2002). Some argue differences do not lie in the text itself[7] but in the reader's framing, echoing reader-response theories (Culler, 2002; Fish, 2003). However, studies have found differences in comprehension (Zwaan, 1991), processing, and affective response (Miall and Kuiken, 1994) of fiction, as well as discourse-level distinctions at scale. Fiction is traditionally associated with narrative immersion and **affective** evocation (Hakemulder, 2020; Scapin et al., 2023; László and Cupchik, 1995), while nonfiction is seen as expository or "indexical", with more explicit, compressed language (Widdowson, 1984; Lehman, 1998; Barth et al., 2022; McIntosh, 1975; Bostian, 1983; Jakobson, 1981). News discourse, for example, tends to be characterized more "disinterested" (Dijk, 2009).

Genre classification studies identify **lexical** and **grammatical** features like adverb/adjective ratios and personal pronouns (Qureshi et al., 2019; Kazmi et al., 2022), type-token ratio (Kubát and Milička, 2013; Sadeghi and Dilmaghani, 2013), nominalization and complexity metrics distinguishing fiction from nonfiction (Vicente et al., 2021), the latter indexing more nouns, nominalizations, and longer words (Dijk, 2009). Other approaches have used model classification or semantic **embeddings** to detect narrative segments in English, demonstrating the value of automated methods and the more semantic dimension for genre classification (Repo, 2024; Laippala et al., 2019). Still, even the "fiction category" remains internally **heterogeneous**: canonical fiction often mirrors nonfiction in complexity (Wu et al., 2024; Bizzoni et al., 2024b), while popular fiction is simpler. Moreover, feuilleton novels in turn have their own distinct characterization: accessible language and emotional pacing, including cliffhangers (Eco, 1967; Lehrmann, 2018; Christoffersen, 2022).

metrics, which models change in literary language to support (re)writing Danish literary history in the long 19th century.

[7]"There is nothing inherently different in the form of literary language" (Stockwell, 2002, p. 7).

3 Data

Collection. The dataset consists of articles from three 19th-century Danish local newspapers[8] – published in Lolland-Falster, Thisted, and Aarhus – digitized as part of the ENO project (see Table 1).[9] To improve OCR quality, particularly for early 19th-century titles, the project uses Transkribus. The output is segmented into articles using a hybrid pipeline, combining rule-based heuristics (e.g., common headers) with a random forest classifier drawing on heterogeneous features such as line length and sentence embeddings. The variation in layout poses additional segmentation challenges.

Selection. In sum, 1,394 articles (i.e., segments) were selected and annotated for their category. These included fiction/nonfiction, as well as some subcategories (see Appendix C). The articles for annotation were in part randomly selected and in part gathered with the intent to locate the serialized novels (batches of fiction and nonfiction articles were collected based on a set of search words, such as "to be continued").

Segmentation. As the newspaper segmentation was prone to errors, especially with long running text (like fiction), feuilleton texts were often split into multiple articles. As the end goal is to classify segmented articles, annotated feuilleton pieces were kept in the same state, but tracked by assigning individual IDs to individual feuilleton series.

	fiction	nonfiction	total
All articles	650	744	1,394
Articles >100 words	413	540	953
Number of series	161		

Table 1: Number of annotated datapoints in each category. Number of raw articles and after filtering, as well as number of full series.

[8][The annotated data will be available upon publication]

[9]Hosted by the Historical Data Lab at Aalborg University: https://hislab.quarto.pub/eno/.

4 Method

4.1 Annotation

Two annotators with backgrounds in literary and religious studies annotated articles for "fiction" and "nonfiction". They classified articles by matching them to a feuilleton series or referencing the article in the scanned newspaper.[10] In ambiguous cases, annotators discussed and assigned specific subcategories (see Appendix C). Of these subcategories, we included 'biography' as part of fiction for its conceptual and narrative similarity.

4.2 Features

4.2.1 Baseline features

MFW100: frequencies of the 100 most frequent words across the dataset, normalized for article length. **TF-IDF**: the text frequency, inverse document frequency of words (max 5,000 words).

4.2.2 Selected features

Feature selection was motivated by previous work to capture key dimensions of literary language (for details, see Appendix D).

Structural complexity. Avg. word and sentence length, dependency distances, and nominal/verb ratio are known proxies for syntactic and surface-level complexity, often considered to be at higher levels in nonfiction (Widdowson, 1984; Jakobson, 1981). Frequencies of 'of' and 'that' further gauge nominal style (Wu et al., 2024).

Stylistic and grammatical profile. We used function word frequencies – powerful stylistic markers (Eder, 2011) – as well as POS-based ratios – personal pronouns, adverb/adjective, and passive/active verbs – known to differentiate fiction and nonfiction (Qureshi et al., 2019).

Lexical features. We computed type-token ratios (overall, nouns, verbs) and a compression ratio to capture lexical richness (Wu et al., 2024).

Affective features. The affective dimension might be more explicit, if not prevalent, in general fiction than nonfiction (Dijk, 2009). Normalized absolute intensity, mean and standard deviation of sentence-level sentiment scores (via `MeMo-BERT-SA`) were used to assess overall sentiment and intra-text sentiment variability (Feldkamp et al., 2025; Bizzoni et al., 2024a).[11] Four models were tested to select `MeMo-BERT-SA`, see Appendix B.

4.2.3 Embeddings

To select embeddings, we defined a benchmarking task, testing six open, non-instruct embedding models (see Appendix A). `jina-embeddings-v3` emerged as the best model for our purposes.[12] We encoded documents, retrieving vectors of 1024 dimensions.[13] 1.5% of texts exceeded the maximum token length and were embedded as the mean of two chunks (see Appendix A).

4.3 Classification model

Preprocessing. We balanced the dataset by undersampling the majority class (nonfiction). Results are reported on the full set and a subset excluding very short texts (<100 words) to observe potential improvements with seleted features (see Table 1).
Model. We used a Random Forest (RF) classifier with 5-fold cross-validation. RFs are robust to overfitting, handle multicollinearity, and can model complex interactions, making them ideal for distinguishing fiction from nonfiction where features may interact in nuanced ways.
Data leakage & overfitting. To prevent data leakage and overfitting on particular feuilleton-series, we ensured that fiction pieces from the same serial narrative never appeared simultaneously in both the training and test sets. We used the `sklearn` implementation of `StratifiedGroupKFold` for this, which aims to preserve class balance in test and training sets while allowing for us to group by feuilleton ID, ensuring that the same feuilleton piece was not split across train and test sets.

5 Results

5.1 Classification: comparing pipeline settings

We present our results in Table 2. Embeddings perform best overall, though the gains over other feature sets are marginal. Notably, TF-IDF alone works as a close runner-up in precision, recall, and F1-scores when compared to embeddings. It is also worth noting that MFW100, TF-IDF, and selected features show improvements on the filtered set (scores in parentheses in Table 2). The discrepancy between recall and precision – with precision

[10] Available via the Danish Royal Library: https://www2.statsbiblioteket.dk/mediestream/

[11] Very long sentences (0.15% of all sentences $n = 19{,}674$) were split into segments due to model input limits.

[12] https://huggingface.co/jinaai/jina-embeddings-v3

[13] The code to retrieve embeddings is available at: https://anonymous.4open.science/r/encode_feuilletons-6922

Features	Class	Precision	Recall	F1-Score
MFW100	*Fiction*	0.84 ± 0.03 (0.87)	0.86 ± 0.03 (0.88)	0.85 ± 0.02 (0.87)
	Nonfiction	0.86 ± 0.02 (0.88)	0.84 ± 0.04 (0.86)	0.85 ± 0.02 (0.87)
TFIDF	*Fiction*	0.84 ± 0.02 (0.86)	0.90 ± 0.01 (0.89)	0.87 ± 0.01 (0.88)
	Nonfiction	0.89 ± 0.01 (0.89)	0.82 ± 0.03 (0.86)	0.86 ± 0.01 (0.87)
Selected features	*Fiction*	0.84 ± 0.03 (0.86)	0.85 ± 0.03 (0.88)	0.84 ± 0.02(0.87)
	Nonfiction	0.85 ± 0.03 (0.88)	0.83 ± 0.04 (0.86)	0.84 ± 0.03 (0.87)
Embeddings	*Fiction*	$\underline{0.88 \pm 0.02}$ (0.89)	$\underline{0.93 \pm 0.01}$ (0.91)	$\underline{0.91 \pm 0.02}$ (0.90)
	Nonfiction	$\underline{0.93 \pm 0.01}$ (0.91)	$\underline{0.88 \pm 0.03}$ (0.89)	$\underline{0.90 \pm 0.02}$ (0.90)

Table 2: Average classification performance over all folds. For each feature set and class: performances on the full dataset and the subset filtered for text length in parenthesis. Highest performance per metric and setting underlined.

higher for nonfiction, and recall higher for fiction – suggests that it is easier to classify nonfiction, possibly due to fiction class heterogeneity.

Considering the effectiveness of function words and lexical frequencies for genre classification, it should be noted that MFW100 and TF-IDF are strong baselines. This makes it all the more impressive that a few selected features can perform nearly as well, reflecting the significant differences in the type of language used in news articles vs. feuilleton novels.

feature	importance
personal pronoun frequency	0.195
nominal/verb ratio	0.114
sentiment intensity	0.089
word length (avg)	0.089
active verb ratio	0.063
passive verb ratio	0.056
sentiment (SD)	0.052
functionword ratio	0.039

Table 3: Avg. feature importances in the RandomForests classifier across 5 folds (top 8 features).

5.2 Modeling fictionality: feature patterns

Beyond performance, we examine linguistic features in fiction vs. nonfiction. Fiction shows greater sentiment variability and more frequent personal pronouns, in line with research linking fiction to immersive, emotive language (Hakemulder, 2020; Zwaan, 1991). Three affective features rank among the top 10 in our selected-features model (see Table 3). Fiction shows both higher sentiment intensity and greater variability in sentiment direction (SD) (see Appendix D, Figure 2). In contrast, nonfiction displays higher information density – reflected in nominal ratio, passive voice, and word length (Fig. 2), also confirming the weight of nouns and nominalizations attributed to nonfiction in Vicente et al. (2021). Function words are especially informative, appearing in both frequency models and feature rankings (Table 8) and feature importance rankings (Table 3). This aligns with stylometric research, highlighting function word frequencies in detecting authorial or genre differences (Eder, 2011; Sobchuk and Šeļa, 2024). Moreover, Qureshi et al. (2019) found that two simple features – adverb/adjective ratio and personal pronoun ratio – are effective in distinguishing modern fiction from nonfiction. In our case, this holds especially for personal pronouns. Complexity measures like dependency length and TTR show limited discriminative power, likely due to the stylistic range of serialized fiction.[14]

6 Discussion & conclusions

Despite the blurred and historically contingent boundary between fiction and nonfiction, our results are promising. Using both embedding-based and feature-based classification, we achieve F1 scores up to 0.91, indicating that linguistic cues – especially affective dynamics and information density – reliably signal fictionality. These findings support two main conclusions: (1) fiction classification is feasible even in noisy, mixed-genre newspaper corpora; and (2) linguistic profiling confirms (some) presuppositions on fiction as a macrogenre. Low-level features and function words are especially strong discriminators, with a model based solely on TF-IDF features performing notably well. Moreover, among interpretable features, information density, surface complexity, and affective features emerge as strong fictionality markers.

[14] Consider that Dickens and Dostoevsky – both canonical authors – serialized their works.

Limitations

The limitations of this study include the relatively narrow temporal scope (1818–1848); future work could extend this range to explore longer-term developments. The analysis is also limited to a small selection of provincial newspapers, deliberately excluding the more widely circulated Copenhagen titles. Although this reflects our focus on noncanonical and locally curated archives, fictionality may manifest differently in more mainstream publications.

Additionally, we use the terms fiction and nonfiction in a broad, categorical sense, even though the fiction treated here, the feuilleton novel, is far from uniform or representative of fiction *tout-court*. Discourse-style distinctions may not align neatly with contemporary notions of fictionality or literariness. Future work could incorporate genre-sensitive modeling or multi-label classification to reflect these subtleties better.

Acknowledgments

Left blank for review.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

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.

Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment classification of historical Danish and Norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Florian Barth, Hanna Varachkina, Tillmann Dönicke, and Luisa Gödeke. 2022. Levels of Non-Fictionality in Fictional Texts. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 27–32, Marseille, France. European Language Resources Association.

Yuri Bizzoni and Pascale Feldkamp. 2023. Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 219–228, Tokyo, Japan. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024a. Good books are complex matters: Gauging complexity profiles across diverse categories of perceived literary quality. *Preprint*, arXiv:2404.04022.

Yuri Bizzoni, Pascale Feldkamp Moreira, Ida Marie S. Lassen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024b. A matter of perspective: Building a multi-perspective annotated dataset for the study of literary quality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 789–800, Torino, Italia. ELRA and ICCL.

Lloyd R. Bostian. 1983. How active, passive and nominal styles affect readability of science writing. *Journalism quarterly*, 60(4):635–670.

Anna Christoffersen. 2022. "A series of waves" : melodramatic rhythms in Victorian serial fiction.

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

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

Umberto Eco. 1967. Rhetoric and ideology in Sue's "Les mystères de Paris". *International Social Science Journal*, 4(19):551–569.

Maciej Eder. 2011. Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint. *Studies in Polish Linguistics*, Volume 6 (2011)(Vol. 6, Issue 1):99–114.

Pascale Feldkamp, Márton Kardos, Kristoffer Nielbo, and Yuri Bizzoni. 2025. Modeling multilayered complexity in literary texts. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 142–158, Tallinn, Estonia. University of Tartu Library.

Pascale Feldkamp, Jan Kostkan, Ea Overgaard, Mia Jacobsen, and Yuri Bizzoni. 2024a. Comparing tools for sentiment analysis of Danish literature from hymns to fairy tales: Low-resource language and domain challenges. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 186–199, Bangkok, Thailand. Association for Computational Linguistics.

Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig, and Kristoffer Nielbo. 2024b. 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.

Pascale Feldkamp, Ea Lindhardt Overgaard, Kristoffer Laigaard Nielbo, and Yuri Bizzoni. 2024c. Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond. In *Computaitonal Humanities Research 2024*. CEUR Workshop Proceedings.

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

Frank Hakemulder. 2020. Finding Meaning Through Literature. *Anglistik*, 31(1):91–110. Publisher: Universitätsverlag WINTER GmbH Heidelberg.

Hans Hertel. 2018. *Den daglige bog: bøger, formidlere og læsere i Danmark gennem 500 år*. Lindhardt og Ringhof.

Eric Heyne. 2001. Where Fiction Meets Nonfiction: Mapping a Rough Terrain. *Narrative*, 9(3):322–333. Publisher: Ohio State University Press.

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

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.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13.

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.

Daniel W. Lehman. 1998. *Matters of Fact: Reading Nonfiction Over the Edge*, 1st edition edition. Ohio State University Press, Columbus.

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

Lei Lei and Matthew L. Jockers. 2020. Normalized Dependency Distance: Proposing a New Measure. *Journal of Quantitative Linguistics*. Publisher: Routledge.

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

J. László and Gerald Cupchik. 1995. The role of affective processes in reading time and time experience during literary reception. *Empirical Studies of the Arts*, 13:25–37.

Carey McIntosh. 1975. Quantities of qualities: Nominal style and the novel. *Studies in Eighteenth-Century Culture*, 4(1):139–153.

David S. Miall and Don Kuiken. 1994. Foregrounding, defamiliarization, and affect: Response to literary stories. *Poetics*, 22(5):389–407.

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

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.

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.

Giulia Scapin, Cristina Loi, Frank Hakemulder, Katalin Bálint, and Elly Konijn. 2023. The role of processing foregrounding in empathic reactions in literary reading. *Discourse Processes*, 60(4-5):273–293. Publisher: Routledge _eprint: https://doi.org/10.1080/0163853X.2023.2198813.

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

Oleg Sobchuk and Artjoms Šeļa. 2024. Computational thematics: comparing algorithms for clustering the genres of literary fiction. *Humanities and Social Sciences Communications*, 11(1):1–12. Publisher: Palgrave.

Sanja Stajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What Can Readability Measures Really Tell Us About Text Complexity? In *Proceedings of Workshop on natural language processing for improving textual accessibility*, pages 14–22, Istanbul, Turkey. Association for Computational Linguistics.

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.

H. G. Widdowson. 1984. *Explorations in Applied Linguistics*. Oxford University Press. Google-Books-ID: WLpoAAAAIAAJ.

Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. Perplexing canon: A study on GPT-based perplexity of canonical and non-canonical literary works. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

Rolf A. Zwaan. 1991. Some parameters of literary and news comprehension: Effects of discourse-type perspective on reading rate and surface structure representation. *Poetics*, 20(2):139–156.

A Embeddings benchmark

We tested four of the best-performing models on the Massive Text Embedding Benchmark (MTEB)[15] – with the criteria: non-instruct and opensource. We also included the `MeMo-BERT-03` model, which has shown promise for working with Danish historical fiction (Feldkamp et al., 2024b; Al-Laith et al., 2024), as well as the `Old_News_Segmentation_SBERT_V0` model which was used for segmentation of the newspaper corpus used in this study.[16] Complete model names are in Table 4.

To assess the quality of our document embeddings, we defined a clustering-based benchmarking task using our labeled corpus of serialized fiction texts (feuilletons) and nonfiction.

Each article in our dataset is associated with a `feuilleton ID` indicating the serial narrative it belongs to. We loaded precomputed pooled sentence embeddings from the six models, grouping each feuilleton text with its corresponding `feuilleton ID`. Nonfiction texts and those without a feuilleton ID were excluded, ensuring that only serialized texts were included in the dataset.

We then applied k-means clustering to these embeddings,[17] treating it as an unsupervised method to group texts that belong to the same feuilleton. The rationale for this task was to evaluate how well the embeddings capture narrative coherence, stylistic features, and textual similarity within serialized fiction. Specifically, we sought to assess whether the embeddings reflect the internal narrative and stylistic relationships (we suppose to exist) within each feuilleton.

We set the number of clusters k to the number of unique feuilleton IDs in the data ($k = 161$) and compared the predicted clusters against the ground-truth feuilleton groupings using two clustering metrics: Adjusted Rand Index (ARI) and v-measure (V). The resulting scores, presented in Table 5, provide an interpretable measure of how well the embedding space captures narrative similarity.

With `jina-embeddings-v3` outperforming

[15] We picked the Scandinavian subset and removed two of the incomplete tasks: `DKhate` and `DanFeverRetrieval`: https://huggingface.co/spaces/mteb/leaderboard

[16] Note that this model was fine-tuned on pairwise sentence similarity with labels with a newspaper article segmentation task in mind.

[17] We used the `Sci-kit learn` implementation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Model	Source
`bilingual-embedding-large`	https://huggingface.co/Lajavaness/bilingual-embedding-large
`Solon-embeddings-large-0.1`	https://huggingface.co/OrdalieTech/Solon-embeddings-large-0.1
`multilingual-e5-large`	https://huggingface.co/intfloat/multilingual-e5-large
`jina-embeddings-v3`	https://huggingface.co/jinaai/jina-embeddings-v3
`MeMo-BERT-03`	https://huggingface.co/MiMe-MeMo/MeMo-BERT-03
`Old_News_Segmentation_SBERT_V0`	https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0

Table 4: Full model names and urls. Models are ordered by score in MTEB (descending). The `MeMo-BERT-03` model was added to the list for its use in Danish literary studies.

Model	ARI	V
`jina-embeddings-v3`	**0.249**	**0.792**
`bilingual-embedding-large`	0.164	0.702
`Old_News_Segmentation_SBERT_V0`	0.07	0.682
`Solon-embeddings-large-0.1`	0.124	0.681
`multilingual-e5-large`	0.122	0.672
`MeMo-BERT-03`	0.107	0.665

Table 5: Clustering performance of different embedding models on feuilleton article groupings. The V-measure captures the homogeneity and completeness of the clusters; ARI (Adjusted Rand Index) measures the similarity between the predicted clusters and the ground truth, adjusted for chance. The table is ordered by descending v-score, with the highest scores in bold.

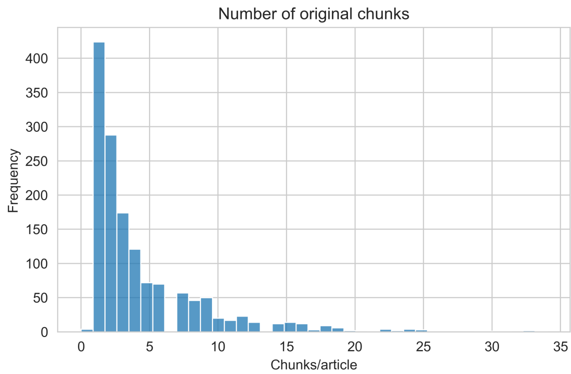


Figure 1: Number of original chunks of articles' embeddings.

other models for this task, we chose this model for our classification of fiction and nonfiction in this study. It is interesting to note that the `Old_News_Segmentation_SBERT_V0` model captures some meaningful structure (good V), but not the precise feuilleton structure (low ARI). This makes it interesting for soft clustering or thematic exploration, but less useful for exact serialized group identification, which is the goal here.

A.1 Pooling embeddings

For all models except `jina-embeddings-v3`, the maximum input length was limited to 514 tokens. In these cases, each feuilleton text was split into chunks of up to 514 tokens, and a mean embedding was computed by averaging across the resulting chunk embeddings. The `jina-embeddings-v3` model, by contrast, supports much longer inputs (up to 8,194 tokens). Only 23 texts exceeded this limit and required splitting into two chunks. For a detailed distribution of the number of chunks required when using models with the 514-token limit, see Fig. 1. Since `jina-embeddings-v3` achieves the highest performance in the clustering task, we suspect that averaging across chunks may dilute meaningful semantic signals, potentially reducing clustering quality.

B Sentiment Analysis benchmark

To select an appropriate sentiment analysis method for Danish literary texts from the 19$^{\text{th}}$ century, we evaluated several recent models using benchmark results from Feldkamp et al. (2024a), which compared dictionary-based and transformer-based approaches against human sentiment annotations of literary sentences. For this purpose, we used the `Fiction4Sentiment` dataset[18], an extended version of the dataset used in Feldkamp et al. (2024a).

`Fiction4Sentiment` includes annotated sentences ($n = 6,300$) from English- (1952–1965) and Danish-language fiction (1798–1873), covering a broad range of genres including prose, hymns, and poetry. The dataset is well-suited to our task for three reasons: (1) it is bilingual, allowing for cross-linguistic comparisons; (2) it spans diverse literary genres, aligning with the possible heterogeneity of fiction in our corpus; and (3) its Danish component closely matches the time period of our feuilleton texts, offering a historically proximate and genre-relevant testbed for model evaluation.

We tested 4 transformer-based models

[18]For details on the dataset, see Feldkamp et al. (2024c). Available at: https://huggingface.co/datasets/chcaa/fiction4sentiment

Model	Multilingual	Danish set	English	Da-En translated set
vader (baseline)	-	-	0.510	0.544
twitter_xlm_roberta (benchmark)	0.553	0.514	0.596	0.571
xlm-roberta-base-sentiment-multilingual	**0.603**	0.603	**0.610**	**0.592**
danish-sentiment	0.539	0.485	0.595	0.569
da-sentiment-base	0.228	0.447	0.129	0.091
MeMo-BERT-SA	0.465	**0.651**	0.254	0.256

Table 6: Spearman correlations of sentiment models' scores with the human gold standard. Columns from left to right: Overall evaluation on English and Danish `Fiction4Sentiment` sentences ($n = 6,300$), evaluation of the Danish subset of sentences ($n = 2,800$), as well as overall evaluation on the Dataset in English, where Danish sentences were translated. Evaluation of the translated set (Da-En) shown in the last right-hand column. Rows from top to bottom: The first two rows are the baseline – VADER (only on English) – and the benchmark on this dataset from Feldkamp et al. (2024a). The best model performance per Dataset setting is in bold, and the follow-up is underlined. Note: All p-values < 0.01.

as well as a dictionary-based method as a baseline. We also included the model to beat from Feldkamp et al. (2024a), i.e., the `twitter-xlm-roberta-base-sentiment`. These were:

VADER,[19] a dictionary-based approach, which we presently use as a baseline.

`twitter-xlm-roberta-base-sentiment`, which was the best performing model in Feldkamp et al. (2024a);[20]

`xlm-roberta-base-sentiment-multilingual`, a finetuned model of the previous, chosen for being multilingual and widely used across languages;[21]

`da-sentiment-base`,[22] based on the aforementioned `twitter-xlm` and fine-tuned on Danish. The model performed best in a binary sentiment classification benchmark in Allaith et al. (2023);

`da-base-sentiment` chosen for being recent and included in the recent benchmark for binary classification (Allaith et al., 2023);[23]

`MeMo-BERT-SA`, a model finetuned for SA on sentences of 19th century Danish novels.[24]

Each model was applied to score sentences against a gold standard. Like Feldkamp et al. (2024c), we used the model confidence score to convert binary model labels (positive, negative) to a continuous score (between -1 through neutral – 0 – to 1), i.e., to scale it like the human judgements. For more on this approach, see Feldkamp et al. (2024a); Bizzoni and Feldkamp (2023). To test the models, we also included scoring on Danish sentences that were translated via Google translate.[25] We did this because Feldkamp et al. (2024a) found that models applied to translated sentences were outperforming the same models applied to the original (Danish) language.

Results are shown in Table 6. Even if we find that `xlm-roberta-base-sentiment-multilingual` performs consistently well across all settings, the `MeMo-BERT-SA` model performs the best on Danish – beating the baseline of Feldkamp et al. (2024a) – which is why we use it for SA in this study.[26]

C Annotation Scheme

Label	Count	Modified
Nonfiction	688	744
Fiction	517	650
Biography	133	fiction
Anecdote	51	remove
Essay	46	nonfiction
Poem	14	remove
Speech	10	nonfiction

Table 7: Distribution of annotated genres in the corpus and modifications for the fiction/nonfiction binary classification.

Fiction was further divided into 'biography', 'anecdote', and 'poem', while 'essay' and 'speech' were subdivisions of nonfiction. Anecdotes and

[19] https://github.com/cjhutto/vaderSentiment

[20] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[21] https://huggingface.co/cardiffnlp/xlm-roberta-base-sentiment-multilingual

[22] https://huggingface.co/vesteinn/danish_sentiment

[23] https://huggingface.co/alexandrainst/da-sentiment-base

[24] https://huggingface.co/MiMe-MeMo/MeMo-BERT-SA

[25] We used the python implementation googletrans: https://pypi.org/project/googletrans/

[26] The full code for replicating this sentiment analysis benchmark is available at: https://anonymous.4open.science/r/literary_sentiment_benchmarking-CF00

poems were excluded from the fiction category due to their brief length and distinct tone. Essays and speeches were similarly excluded from nonfiction for their narrative structures. See table 7. A detailed table of the annotation scheme and instructions is found in the repository accompanying this paper: https://anonymous.4open.science/r/factfiction_newspapers-E174.

D Features

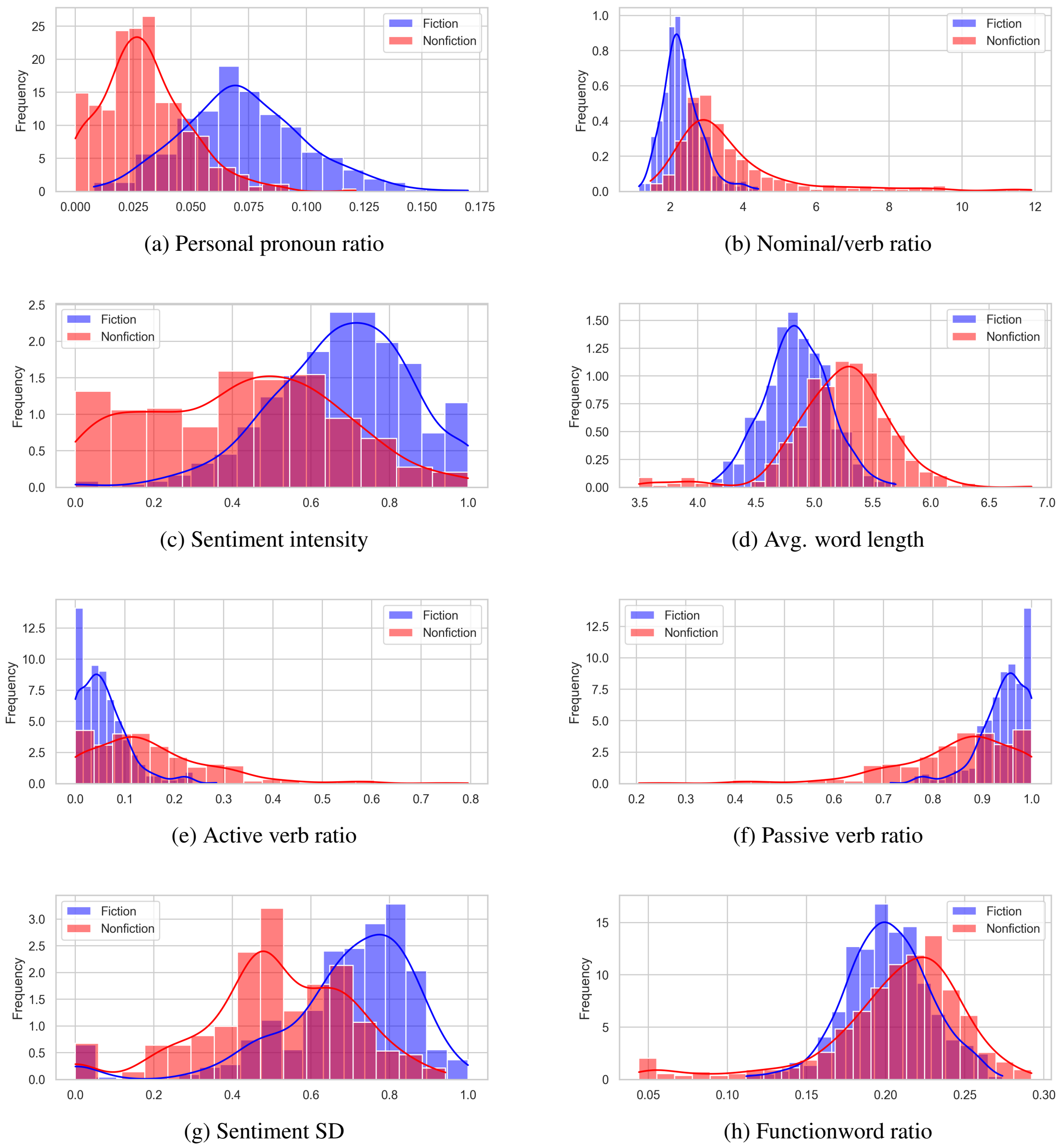
D.1 Feature importances, MFW100

D.2 Feature differences, fiction/nonfiction

D.3 Selected features

word	importance
han	0.064
jeg	0.055
ham	0.055
var	0.037
mig	0.030
de	0.029
skal	0.026
af	0.025
har	0.024
hans	0.020
hun	0.018
er	0.018
havde	0.018
fra	0.018
sagde	0.017

Table 8: Avg. feature importances – top 15 most important words (of the MFW100) – of the RandomForests classifier across 5 folds. Note that importances (all 100 words) sum to 1.



(a) Personal pronoun ratio

(b) Nominal/verb ratio

(c) Sentiment intensity

(d) Avg. word length

(e) Active verb ratio

(f) Passive verb ratio

(g) Sentiment SD

(h) Functionword ratio

Figure 2: Difference in feature levels between fiction and nonfiction groups in the top 8 features in feature importance for the classification (over 5 folds), see table 3. Note that the very short texts (<100 words) were dropped in these plots. For all of these distributions, a t-test shows a significant difference between fiction and nonfiction.

Type	Feature	Description
Surface- and structure-level complexity	**Word and sentence-length**	Longer words and sentences are frequently used in more formal or complex registers, indicate increased cognitive load for the reader, and are frequently used in readability formulae (Stajner et al., 2012). Used for fiction/nonfiction classification in Kazmi et al. (2022).
	Normalized Dependency Distance, mean & SD	Quantifies the mean and SD in dependency length as indicators of structural complexity in texts. We followed the procedure for normalization proposed in Lei and Jockers (2020).
	Nominal verb ratio	Quantifies the proportion of nouns and adverbs (over verbs) in the text, reflecting the nominal tendency in style, which is often associated with complex linguistic structures, denser communicative code, expert-to-expert communication (McIntosh, 1975; Bostian, 1983). The predominance of nouns and nominalizations was found to be important for distinguishing news articles in Vicente et al. (2021).
	"Of"/"that" frequencies	Frequency of these function words have been seen to indicate, in the case of "of", a more nominal prose, and in the case of "that", a more declarative and verb-centered prose. Wu et al. (2024)
Stylistic and grammatical profile	**Function words**	Frequency of function words (normalized for text length), suggesting a more information-rich prose when lower.
	Personal pronoun ratio	Proposed as a strong fiction/nonfiction marker in Qureshi et al. (2019).
	Averb/Adjective ratio	Proposed as a strong fiction/nonfiction marker in Qureshi et al. (2019)
	Passive and active verb ratio	Heigthened use of passive verbs can suggest structural complexity and more nominal styles (Bostian, 1983).
Lexical features	**Type-Token Ratio (MSTTR-100)**	Measures lexical diversity by comparing the variety of words (types) to the total number of words (tokens), indicating a text's vocabulary complexity and inner diversity. A high TTR represents a richer prose: a higher diversity of elements and a lower lexical redundancy (**?**). We used the Mean Segmental Type-Token Ratio (MSTTR). MSTTR-100 represents the overall average of the local averages of 100-word segments of each text. Diversity was used to differentiate between genres (Sadeghi and Dilmaghani, 2013) and MSTTR specifically was used to classify fiction/nonfiction (Kazmi et al., 2022).
	TTR Noun, TTR Verb	TTR of nouns or verbs quantifies the same diversity as above within these Parts-of-Speech categories. Nouns and verb variability is correlated with more demanding prose (Wu et al., 2024).
	Compressibility	Measures the extent to which the text can be compressed, serving as an indirect indicator of redundancy and lexical variety (**?**). We calculated the compression ratio (original bit-size/compressed bit-size) for the first 1500 sentences of each text using bzip2, a standard file-compressor, as in Koolen et al. (2020).
Affective features	**Sentiment intensity, mean & SD**	Represents the intensity (absolute value), average and variability in sentiment. Sentiment variability has been linked to extended text processing time and perceived difficulty (Feldkamp et al., 2025).

Table 9: Selected features related to stylistic, structural and sentiment complexity and variability.