Why Novels (Don't) Break Through: Dynamics of Canonicity in the Danish Modern Breakthrough (1870–1900)

Alie Lassche[1], Pascale Feldkamp[1], Yuri Bizzoni[1], Katrine Baunvig[2] and **Kristoffer Nielbo[1]**
[1] Center for Humanities Computing, Aarhus University
[2] Center for Grundtvig Studies, Aarhus University
{a.w.lassche, pascale.feldkamp, yuri.bizzoni, baunvig, kln}@cas.au.dk

Abstract

Recent studies suggest that canonical works possess unique textual profiles, often tied to innovation and higher cognitive demands. However, recent work on Danish 19th century literary novels has shown that some non-canonical works shared similar textual qualities with canonical works, underscoring the role of text-extrinsic factors in shaping canonicity. The present study examines the same corpus (more than 800 Danish novels from the Modern Breakthrough era (1870–1900)) to explore socio-economic and institutional factors, as well as demographic features, specifically, book prices, publishers, and the author's nationality – in determining canonical status. We combine expert-based and national definitions of canon to set up a classification experiment to test the predictive power of these external features, and to understand how they relate to that of text-intrinsic features. We show that the canonization process is influenced by external factors – such as publisher and nationality – but that text-intrinsic features nevertheless maintain predictive power in a dynamic interplay of text and context. To ensure reproducibility, code and raw data are available at https://github.com/centre-for-humanities-computing/text-extrinsic-canon.

1 Introduction

Why do some novels have an enduring status in literary cultures while others remain outside the canon? The question of how novels achieve – or fail to achieve – canonical status has long fascinated literary scholars, generating a rich field of study. Recent work suggests that the textual features of literary works hold significant predictive power in determining their canonicity. Compared to non-canonical works, canonical works exhibit a unique textual profile (Barré et al., 2023; Brottrager et al., 2021; Porter, 2018), with stylistic characteristics connected to a higher cognitive load on the reader (Bizzoni et al., 2024; Wu, 2023; Wu et al., 2024).

Moreover, recent studies have gone beyond stylistic analysis to examine representations of canonical novels in semantic space. For example, Barré (2024), working with a corpus of historical French fiction, demonstrated that canonical works are often more deeply integrated into an intertextual network after publication. Similarly, in Feldkamp et al. (2024b) we examined textual embeddings of late 19th-century Danish novels, revealing that canonical novels distinguish themselves through innovation and impact. These novels not only stand out from their contemporaries but also appear to be literary trendsetters of their time.

Although previous studies have reaffirmed the role of textual features in determining a novel's canonicity, they do not fully explain the phenomenon. Either the features selected for analysis or the definition of the "canon" appear to create blind spots. For instance, in Feldkamp et al. (2024b) we identified a category of novels with textual profiles similar to canonical works, which, however, remain lesser known today. This suggests that textual qualities alone may not be sufficient to explain canonicity. The inability of these "non-canonical canonicals" (i.e., novels with textual profiles similar to canonical works) to achieve widespread recognition implies that other factors – beyond the textual features – play a crucial role in shaping canonicity.

Previous research has emphasized the importance of text-extrinsic factors such as the spread of novels, their accessibility to readers, and the socio-economic conditions surrounding their production (Heydebrand and Winko, 1996; Guillory, 1995). These aspects may influence canonization processes, where evaluation plays a role at every level, from publisher to reviewer and reader (Heydebrand and Winko, 1996; Brottrager et al., 2021) and where institutions also create and maintain the

Proceedings of LaTeCH-CLfL 2025, pages 278–290

canon (Guillory, 1995). Such factors may be key to understanding why some works with seemingly "canonical" characteristics fail to enter the canon.

Thus, the case of Feldkamp et al. (2024b)'s non-canonical novels raises an important question: are models which focus primarily on text-intrinsic features overlooking key factors related to a novel's dissemination and reception? To answer this question, we investigate the broader socio-economic and institutional contexts of literary production, focusing on text-extrinsic factors – specifically, book prices, publishing houses, and the author's nationality – as predictors for a novel's canonicity.

We test the strength of text-extrinsic features for determining the canonical status of a novel in a classification task. We compare this to the performance of exclusively text-intrinsic features as used in Feldkamp et al. (2024b), as well as the combination of text-extrinsic features and text-intrinsic features. We propose two hypotheses:

H1: Novels that achieve canonical status are more strongly associated with a combination of text-intrinsic and text-extrinsic features (than with, e.g., text-intrinsic features alone).

H2: Novels that achieve canonical status are more strongly associated with either text-intrinsic or text-extrinsic features (such that the addition of, e.g., text-extrinsic features does not significantly improve the prediction of canonicity).

Our classification task with different text-intrinsic and text-extrinsic settings will give us an idea of how these factors interact in literary canon formation. Moreover, we inspect models based on all possible feature combinations individually and analyse misclassifications in depth to gauge what they can tell us about the boundaries of the literary canon.

For this study, we use the same corpus of novels from the Modern Breakthrough (*det Moderne Gennembrud*, 1870-1900) as we did in Feldkamp et al. (2024b), to examine them in a controlled context. This period is ideal for our study because it offers exhaustive coverage of literary production within a short timeframe, situating the novels within a small, relatively contained literary field (the Danish). This approach is significant because previous efforts to examine canonicity often struggle to account for the "dark numbers" of literary production – i.e., the forgotten or "great unread" works (Moretti, 2000). By focusing on a small, restricted, yet exhaustive setting, we can directly compare canonical novels to the contemporary production, avoiding the potential biases introduced by spuriously selected control groups.

This paper is structured as follows: Section 2 reviews related work on text-intrinsic and text-extrinsic features of canonical works, as well as the literary context of our corpus. Section 3 provides an overview of the corpus used in this study and explains how the canonicity of a novel was defined. Section 4 details our methodological pipeline, covering the creation of document representations, selection of text-extrinsic features, preparation of classification tasks, execution of experiments, and analysis of false positives. Section 5 presents the results, beginning with descriptive statistics, followed by the classification outcomes and an in-depth analysis of false positives. This is followed by a discussion in Section 6, and concluding remarks in Section 7.

2 Related Work

2.1 Features of the canon

The discussion about canon has often focused on the tension between two perspectives: one that views canonicity as conferred "from above", based on cultural, political, or institutional factors (Guillory, 1995), and another that sees it as a reflection of the inherent excellence of the works "from below", grounded in text-intrinsic features (Bloom, 1995). Recent studies have offered a more nuanced view of this debate. They demonstrate that text-extrinsic features[1] are strong predictors of canonicity (Brottrager et al., 2021), but also confirm that canonical works possess distinctive text-intrinsic characteristics compared to non-canonical works (Feldkamp et al., 2024b; Barré et al., 2023; Brottrager et al., 2021; Porter, 2018). Furthermore, canonical works exhibit textual profiles that differ not only from non-canonical works but also from other categories of literary recognition, such as bestselling or prize-winning novels (Bizzoni et al., 2024; Wu et al., 2024). For distinguishing canonical works on the large scale, studies have mainly focused on stylistic or syntactic features (Algee-Hewitt et al., 2016; Brottrager et al., 2021), such as linguistic measures related to a novel's complexity (Wu et al., 2024). Notably, this has been an attempt to gauge stylistic/syntactic differences between canon and non-canon overall, and not within a given field or period. As such, the more contex-

[1] I.e., cultural, political, or market traits, as in Wang et al. (2019).

tual, but also the semantic aspects of literary texts have been relatively overlooked. Still, recent studies like Barré (2024) use text embeddings to show how canonical works appear to have a stronger echo in the literary field after their publication than non-canonical works have – perhaps a stronger presence in shaping norms and trends for literature, which can here be interpreted as semantic.

Considerable recent work has examined indicators of canonicity, shedding light on their interrelations (Brottrager et al., 2022, 2021; Feldkamp et al., 2024a; Barré et al., 2023; Algee-Hewitt et al., 2016). For instance, school-based and scholarly indicators of canonicity appear more closely linked, while prize lists tend to be more disparate, revealing a complex interplay of actors in canonization (Barré et al., 2023; Feldkamp et al., 2024b).

However, little data-driven research has investigated how a work's canonization relates to factors of literary production in its historical context, such as the role of its publishing house. Prominently, Winko (2002) describes canonization as an emergent process shaped by numerous uncoordinated yet intentional actions, where individual choices accumulate over time. While some actors, such as institutions, play a more influential role as guardians or shapers of the canon, the impact of different actor types remains conjectural, and even recent studies question the role of text-intrinsic features (Herrmann, 2011).[2]

Building on this literature, the present study firstly tests the relative influence of text-intrinsic features in the process of canonization. Secondly, it compares and examines how specific aspects of the literary system – particularly the role of publishers, accessibility (e.g., prices), and author profile (nationality) – shape the canonization of a work within its historical context.

2.2 The Danish Modern Breakthrough

The Modern Breakthrough was a transformative period in Danish literature, marking the shift from romanticism to realism and naturalism. Spearheaded by Georg Brandes,[3] the movement emphasized literature's role in societal critique, focusing on social issues, individualism, and science (D'Amico, 2016; Bjerring-Hansen and Wilkens, 2023).

At the same time, literary tastes shifted: realist novels rose to prominence, while historical novels, like those by B.S. Ingemann, lost their earlier popularity (Bjerring-Hansen and Rasmussen, 2023; Martinsen, 2012). This polarization between realist and historical literature highlights the evolving dynamics of literary authority, market forces, and reader reception. Realist novels gained a place in the literary canon, while genres like the historical novel declined (Bjerring-Hansen and Wilkens, 2023). Canonicity, therefore, may have been shaped by more than just textual qualities; socio-economic factors, market dynamics, and reader demographics also seem to have played a significant role.

Overall, the Modern Breakthrough was composed by three interdependent shifts: one in literary production (subject and volume of printed literature), one in the literary field (rise and fall of publishers), and one in literary culture (changing reader tastes and demand for accessible literature). The Modern Breakthrough likely led to the rise of certain textual profiles and a more heterogeneous corpus, reflecting the dominance of Realism. Moreover, changes in publishing dynamics and reader preferences may complicate the modeling of the canon. This also means that the period of the Modern Breakthrough, though relatively short in duration (30 years), is anything but a minor period in terms of complexity.

3 Data

Our dataset comprises 838 original Danish and Norwegian novels published between 1870 and 1900, accompanied by metadata such as page count, (original) book price and publishing house. All novels, including those by Norwegian authors, were published in Danish and by Danish publishers. The corpus includes all first-edition novels from Danish publishers during this period, excluding non-novel works like short story collections.[4]

We use the categorization of novels' canonical status in Feldkamp et al. (2024b).[5] Their list of

[2] Herrmann (2011) argues that the idea of textual factors influencing all forms of canon formation is an implicit assumption, neither empirically proven nor accounted for in theoretical descriptions.

[3] Brandes' Copenhagen lecture (1871) and J.P. Jacobsen's *Mogens* (1872) are often considered the start of the Modern Breakthrough (Bjerring-Hansen and Rasmussen, 2023).

[4] This compilation – the MiMe-MeMo corpus – was developed by J. Bjerring-Hansen, P. Diderichsen, D. Haltrup, and N.E.D. Jørgensen, based on the Danish book index. For details, see Bjerring-Hansen et al. (2022). Version 1.1, utilized in this study, is accessible at: `https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1`.

[5] Note that the categorization in Feldkamp et al. (2024b) is author-based, meaning that all books in the corpus by an author mentioned in their canon-list are tagged as canonical, even if it is not the author's most prominent work.

	titles	*authors*
Corpus	838	361
Canon	114	20
Other	724	342

Table 1: Statistics on the corpus.

the canon included authors indexed in the Educational Canon (*Undervisningskanon*) and the Cultural Canon (*Kulturkanon*), introduced by the Danish government in the early 21st century to promote Danish literature and standardize school curricula (Harbild et al., 2004). However, the government-defined canons exclude Norwegian authors and are likely driven by political agendas. To provide a more expert-driven perspective, Feldkamp et al. (2024b) collected a canon list from on the encyclopedia *Den Store Danske*, specifically its entry on 'det moderne gennembruds litteratur.'[6] Novels featured in the Cultural Canon, written by authors mentioned in the Educational Canon, or listed in the entry of 'det moderne gennembruds litteratur' in *Den Store Danske* are labeled as *Canon*, while all others are categorized as *Other*. (See corpus statistics and category details in Table 1.[7])

4 Methods

To test our hypotheses, we take the following approach in this paper:

1. Creating document representations. To build a compact representation of the texts, we use a large language model to create a semantic embedding of each novel – the `m-e5-large-instruct` model.[8] Previous work has tested this model against three other SOTA models for Danish in creating embeddings that would perform well generally and across historical Danish documents for this particular corpus (Feldkamp et al., 2024b).[9] Each novel is divided into chunks of the same size,[10] and embeddings were created for every chunk. The average embedding of all chunks of a novel is then used as a representative embedding for that novel.

2. Selection of text-extrinsic features. For each novel, we collect its first edition price, the editor that published it, and the nationality of the author, to represent some aspects of the novels' text-extrinsic profile. Price and editor could be causes of a novel's canonization, or consequences of the very qualities that ensured its canonization. Nationality, on the other hand, can only act as 'cause' in the selection pattern. We selected these features as a starting point because we expected them to have the strongest impact on canonization and because they exhibit a reasonable distribution across the two classes. Other relevant elements, like reprint history, had to be excluded due to data availability.

3. Preparing classification tasks. We perform a classification task using a simple Random Forest model. Random Forest was chosen for this task because it shows robust performance with mixed data types (continuous and categorical) and, through its ensembling, effectively mitigates overfitting. It is also a robust model, well suited for handling outliers. The two classes we are working with are *Canon* and *Other*.

4. Sampling. Because our two classes are unbalanced, we randomly downsample the larger class (*Other*). In order to guarantee robustness, we repeat the majority class downsampling (and training/testing) 50 times and take the average precision, recall, and F1-score across all 50 runs as our results. In each run, we reserve 10% of the data for testing.

5. Experiments. First, we perform a baseline task in which we use the average sentence length as a feature, we assume this to be a relatively simplistic representation of a novel text. Second, to model the impact of text-intrinsic and text-extrinsic features on the process of canonization, we experiment with the following features: (1) text-intrinsic features, i.e., embeddings, and (2) text-extrinsic features, i.e., price, publisher, and nationality. We run experiments with all possible combinations of these four features.

6. False positives analysis. To detect non-canonical novels that contain a textual profile similar to canonical novels, we closely analyse the false positives from the experiments that result from run-

[6] See `https://denstoredanske.lex.dk/det_moderne_gennembruds_litteratur`. Note that while government canons and *Den Store Danske* index various genres,this paper focuses solely on novels.

[7] An extended dataset with additional tags is available on `https://huggingface.co/datasets/chcaa/memo-canonical-novels`.

[8] `https://huggingface.co/intfloat/multilingual-e5-large-instruct`.

[9] The four models that were tested included the historical Danish `MeMo-BERT` model (Al-Laith et al., 2024), the best-performing Danish sentence encoder `DFM-large` (Enevoldsen et al., 2023), and the two best-performing open-weight models on SEB, `m-e5-large` and its prompt based version `m-e5-large-instruct` (Wang et al., 2024). For a detailed description of the task, see Appendices F and G in Feldkamp et al. (2024b).

[10] Since the maximum chunk size includes the length of the prompt, we use a chunk size of 512 - 87 = 425 characters.

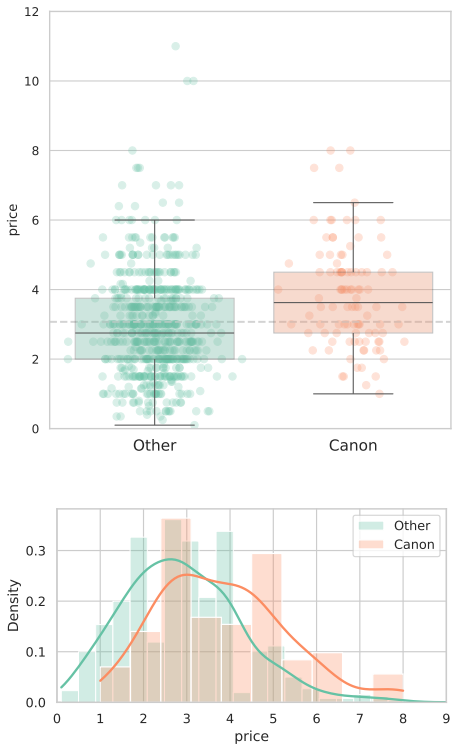

Figure 1: Boxplot (upper) and distribution (KDE) plot (lower) of book price across categories.

ning the model only on text-intrinsic features, i.e., embeddings. That is, we are interested in false positives where *Other* books were misclassified as *Canon*. In order to secure robustness of our results, we run enough iterations to obtain 12 predictions for each novel.[11]

5 Results

5.1 Descriptive statistics

Inspecting the descriptive statistics, we find that both the distribution of publishing houses and original book price vary between categories. The upper plot in Figure 1 shows the distribution of book prices per label category, depicted in two boxplots. The bottom figure shows a kernel density estimate (KDE) plot of each label category. On average, prices of canonical books are higher than those of non-canonical books. The heatmap in Figure 3 in Appendix A shows that almost all canonical books are published by a handful of the largest publishing houses (Gyldendal, Reitzel, Schubothe, Det Nordiske Forlag, Schou, and Philipsen). Together,

[11]In other words, we run enough iterations of our embedding-based classification model to ensure that every ‘ *Other* book is included in a test-set 12 times.

	Canon		Other	
	titles	*authors*	*titles*	*authors*
Danish	68%	70%	86%	89%
Norwegian	32%	30%	13%	10%
German	0%	0%	1%	1%

Table 2: Distribution of author nationalities within the corpus, based on number of authors and novels.

these six publishing houses are responsible for 94% (107) of all canonical novels. However, this does not immediately imply that the larger the publishing house, the higher the chance a novel becomes canonical. There are other large publishing houses where no canonical novels were published (Jyds Forlags-forretning and A. Behrend, for example), and smaller publishing houses with a more even canon/non-canon ratio. Furthermore, these statistics show that publishing houses that are responsible for a large part of the canon production, also publish non-canonical books.

In Table 2, we present the distribution of author nationalities within our corpus, including both the distribution of unique authors and the distribution of all novels. Beyond Danish authors, the corpus includes works by Norwegian authors and a few German authors. The proportion of canonical novels written by Norwegian authors is notably higher than in the non-canonical group (32% versus 13%). In our classification tasks, we further examine the influence of the author's nationality on a novel's likelihood of achieving canonical status.

5.2 Classification tasks

The average performances of the classification experiments are summarized in Table 3. In nearly all experiments, the baseline performance based on average sentence length is surpassed. Embeddings alone appear to be strong predictors for canonicity, yielding F1-scores of 0.728 for the *Canon* class and 0.677 for the *Other* class. This aligns with the findings of Feldkamp et al. (2024b), suggesting that canonical novels possess a distinctive textual profile that sets them apart from the broader literary corpus. This result becomes even more impressive when we take into account that we are using an very rough representation of the novels – the texts are reduced to a set of semantic embeddings (of which we cannot say with certainty what exactly they do and do not capture), of which we then take the average.

However, several text-extrinsic features or com-

Type	Feature set	Precision		Recall		F1-score	
		Canon	*Other*	*Canon*	*Other*	*Canon*	*Other*
Baseline	avg_sentence_length	0.511	0.514	0.828	0.213	**0.585**	0.222
Text-extrinsic	price	0.551	0.553	0.560	0.534	**0.545**	0.534
	publisher	0.647	0.864	0.909	0.501	**0.753**	0.620
	nationality	0.633	0.549	0.293	0.839	0.389	**0.662**
	price_publisher	0.648	0.676	0.683	0.622	**0.658**	0.638
	price_nationality	0.580	0.580	0.551	0.601	0.554	**0.581**
	publisher_nationality	0.647	0.857	0.905	0.505	**0.752**	0.624
	price_publisher_nationality	0.657	0.684	0.691	0.637	**0.667**	0.652
Text-intrinsic	embeddings	0.681	0.764	0.795	0.624	**0.728**	0.677
Combination	embeddings_price	0.685	0.754	0.780	0.639	**0.723**	0.683
	embeddings_publisher	0.684	0.738	0.772	0.627	**0.718**	0.667
	embeddings_nationality	0.693	0.764	0.790	0.642	**0.731**	0.686
	embeddings_price_publisher	0.694	0.775	0.804	0.641	**0.739**	0.692
	embeddings_price_nationality	0.688	0.756	0.782	0.642	**0.726**	0.685
	embeddings_publisher_nationality	0.691	0.756	0.783	0.643	**0.728**	0.686
	embeddings_price_publisher_nationality	0.690	0.749	0.778	0.643	**0.726**	0.684

Table 3: Performance of Random Forest models based on a baseline (avg. sentence length) and different feature sets: text-extrinsic features only, text-intrinsic feature (embeddings), and a combination of text-extrinsic and -intrinsic features. The dataset is down-sampled to have balanced classes (114 data points per class). Values represent average results across 50 iterations. In green: the best settings for that class. In **bold**: the best predicted class for those settings.

binations thereof also obtain a high performance when predicting canonicity in our corpus. The average F1-scores range between 0.389 and 0.753. Some of these outperform the text-intrinsic features: the highest performance for the *Canon* class is achieved using the publishing house as the sole feature (0.753), followed closely by the combination of publisher and nationality (0.752). This reveals that text-extrinsic features also serve as good predictors for a novel's inclusion in the canon.

When text-extrinsic features are combined with embeddings, F1-scores for the *Canon* class fall within the range of 0.718 to 0.739, suggesting that together they achieve a similar performance in predicting canonicity. Across experiments – regardless of whether they rely on text-intrinsic or text-extrinsic features – the *Canon* class consistently exhibits better predictive outcomes. Exceptions arise when nationality alone, or in combination with price, are used as features.

To evaluate whether these results are disproportionally influenced by the very long tail of smaller publishing houses – each publishing only one novel – we conduct the same classification experiments on a subset of the corpus. This subset includes novels from the eight publishing houses that each contribute to the dataset with more than 25 novels (see Figure 3). The performance metrics for these experiments are shown in Table 4. Notably, the performance of the text-intrinsic features (embeddings) remain stable, with F1-scores stabilize at 0.713 for both classes. The F1-scores of the text-intrinsic features are slightly lower than in the experiments with the full corpus, and the same goes for the performances in experiments with both text-intrinsic and -extrinsic features. The pattern, observed in Table 3, remains the same: together, these features achieve a similar performance in predicting canonicity as in experiments with only text-intrinsic features.

These findings reinforce the robustness of embeddings in predicting canonicity and suggest that the textual characteristics distinguishing canonical novels are not merely artifacts of data imbalance among publishing houses.

5.3 Not breaking through

As much as these results show us that both text-extrinsic and text-intrinsic features play a role in the process of canonization, they also highlight novels that do not conform to this pattern. The F1-score of 0.728 when using embeddings as feature for the classification task, suggests there are novels with a textual profile similar to canonical works, but that remain lesser known today. In this section, we dive deeper into these false positives (incorrectly classified as *Canon*), to better understand why they failed to achieve canonical status. After predicting each novel 12 times, we filtered for non-canonical novels that were incorrectly labeled as *Canon* when

Figure 2: Ratio of true negatives (TN) and false positives (FP) of non-canonical novels that are incorrectly classified as canonical based on $embeddings$, when using other feature sets. We only include novels of which 75% of the predictions for $embeddings$ are FP. The unique number of novels is 145.

embeddings were used as features. We then applied a second filter, retaining only novels labeled as FP at least 9 times (75%). This resulted in a list of 145 novels, each predicted 12 times.

Figure 2 shows how these 145 novels are predicted when using different sets of features. The stacked bar plots show that when using text-extrinsic features such as publisher, price, nationality, or combinations thereof, the frequency of incorrect predictions for these 145 novels decreases substantially. In other words, text-extrinsic features make it easier to correctly predict these novels as non-canonical compared to embeddings. These false positive novels were published by 32 different publishing houses. Among them, 97 novels (67%) were published by the six houses responsible for most canonical novels: Gyldendal (54), Schubothe (14), Schou (11), Det Nordiske Forlag (9), Reitzel (7), and Philipsen (2). The remaining 48 novels were published by 26 smaller houses, most of which are represented by only one book in our dataset. To explore this phenomenon at the level of individual novels, we created the heatmap in Figure 4 in Appendix B. This visualization includes all 145 false positive novels (as identified using embeddings) and shows how often they were incorrectly classified as canonical when other feature sets were used. A cell value of 1 indicates that the novel was predicted as a false positive in all 12 predictions for that feature set, while a value of 0.5 indicates it was a false positive in 6 out of 12 predictions. Novels are sorted by the sum of their row values (excluding embeddings). The higher a novel appears in this heatmap, the more often our model correctly predicted it as *Other* based on its publisher, price, and the author's nationality.

Two novels that appear prominently in the heatmap (*Forfløjne Pile* and *På Solsiden*) are by Carl Muusmann (1863–1936), a Danish author and journalist who worked for various newspapers, including *Berlingske Tidende* and *Nationaltidende*. Muusmann was particularly known for his crime novels and was considered a pioneer in the genre. The textual style of these two novels is highly similar to canonical works, but they lack the correct combination of publisher, price, and other text-extrinsic features. Interestingly, another of Muusmann's novels, *Bondekunstneren*, appears lower in the heatmap, as its publisher and price align more closely with those of canonical works.

While many non-canonical novels were printed in lower-cost formats, some authors, such as Carl Muusmann, had works produced with considerable material quality. For example, Ilsøe (2014) notes that Muusmann's *Det lille Paradis* (1911) was published by Kunstforlaget Danmark with decorative endpapers designed by Axel Hou, indicating a level of aesthetic investment. This suggests that book material quality alone did not determine canonicity, and that institutional factors may have played a larger role in excluding certain works. Notably, Muusmann never remained with a single publisher; instead, his five novels in the dataset were published by four different houses, suggesting a lack of the institutional backing that often contributes to literary canonicity. Other authors with multiple novels high in the heatmap include Axel Betzonich (*Don Juans Efteraar*, *Peter Jensen*), Jakob Hansen (*Karen Hav*, *Ved Højvande*), and Otto Møller (*Lys over Landet!*, *Overmennesker*, *Millionærens Pilegrimsfærd*). These cases point in the direction of the hypothesis that writing in the correct textual style is insufficient for achieving canonical status; a novel must also have the right publisher, price, and potentially many other contextual attributes.

Conversely, novels at the bottom of the heatmap in Figure 4 exhibit the correct textual profile, publisher, price, and nationality, yet they are still excluded from contemporary canonical lists. This highlights, amongst other things, the inherent limitations of the canon itself. Despite expanding the canon by incorporating expert opinions (e.g., based on *Den Store Danske*), the presence of Norwegian author Jonas Lie in this heatmap (with his novels *Faste Forland*, *Livsslaven*, *Et Samliv*, *Niobe*, *Kommandørens Døttre*, *Thomas Ross*, and *Gå På*) underscores how perceptions of the canon differ

Type	Feature set	Precision		Recall		F1-score	
		Canon	*Other*	*Canon*	*Other*	*Canon*	*Other*
Baseline	avg_sentence_length	0.512	0.667	0.936	0.142	**0.656**	0.200
Text-extrinsic	price	0.536	0.544	0.516	0.564	0.518	**0.546**
	publisher	0.540	0.622	0.729	0.376	**0.599**	0.428
	nationality	0.627	0.537	0.302	0.807	0.393	**0.642**
	price_publisher	0.571	0.605	0.638	0.516	**0.596**	0.543
	price_nationality	0.559	0.547	0.540	0.564	0.538	**0.546**
	publisher_nationality	0.573	0.571	0.576	0.556	**0.562**	0.549
	price_publisher_nationality	0.573	0.588	0.618	0.527	**0.585**	0.542
Text-intrinsic	embeddings	0.719	0.738	0.724	0.709	**0.713**	**0.713**
Combination	embeddings_price	0.698	0.715	0.709	0.685	**0.695**	0.691
	embeddings_publisher	0.692	0.706	0.711	0.671	**0.694**	0.681
	embeddings_nationality	0.722	0.726	0.720	0.711	**0.713**	0.712
	embeddings_price_publisher	0.701	0.714	0.715	0.684	**0.700**	0.691
	embeddings_price_nationality	0.703	0.730	0.735	0.676	**0.710**	0.692
	embeddings_publisher_nationality	0.705	0.715	0.716	0.685	**0.703**	0.692
	embeddings_price_publisher_nationality	0.698	0.737	0.747	0.667	**0.714**	0.692

Table 4: Performance of Random Forest models based on a baseline (avg. sentence length) and different feature sets: text-extrinsic features only, text-intrinsic features (embeddings), and a combination of text-extrinsic and -intrinsic features. **The dataset only includes the novels of large publishing houses of which we have more than 25 novels in our dataset.** We have down-sampled to have balanced classes (107 data points per class). Numbers represent average results across 50 iterations. In green: the best settings for that class. In **bold**: the best predicted class for those settings.

across national boundaries. While Lie holds canonical status in Norway, he is not included in the version of the Danish canon that was used in this study.

6 Discussion

The results of our classification tasks show that both the text-intrinsic features and (a subset of) the text-extrinsic features provide predictive value of canonicity. In our experiments based on the full corpus, the text-extrinsic features outperform the embeddings. This confirms H2. However, when we only look at the performances of the experiments based on the subset of large publishing houses, text-intrinsic features outperform text-extrinsic features, which confirms our H2 in the opposite direction. A combination of both embeddings and nationality, or all features together, result in similar performances. This does not provide strong support for H1, but since a combination of features does not lower the predictive performance either, it is neither a rejection of this hypothesis. Additional experiments are required to be able to either confirm or reject H1.

In sum, the misclassification of novels, as discussed in the previous section, suggests that textual similarity to canonical works alone is not sufficient for inclusion in the canon: the lack of editorial support or limited distribution due to price choices might impact their status since their first publication. The presence of false positives having the 'right profile' in terms of price, editor and nationality, on the other hand, might indicate two different phenomena: *(i)* there are other text-extrinsic features that impact their canonical status, such as institutional support, inclusion in specific literary trends, and so forth; *(ii)* there are some essential text-intrinsic features, not captured by textual embeddings, that could explain their exclusion from the canonical group. Whether Muusmann and the other mentioned authors were excluded from our canonical lists for the first or the second order of reasons is probably a question for a next study.

There are several directions in which future research could develop. Firstly, the definition of canonicity could be refined, for example by using alternative lists, and by replacing categorical labels with a more continuous metric that better accounts for degrees of recognition. Expanding the range of text-extrinsic features could improve our understanding of how text and context interact with each other in the process of canonization. Additionally, a more detailed analysis of false positives – including their commercial success and literary afterlife – would help contextualize these works. One approach to this would be to do a text re-use study and investigate which novels are more often discussed in public debate – either cited in newspapers

or mentioned in the works of influential critics such as Georg Brandes and Søren Kierkegaard. Moreover, it would be equally worthwhile to investigate the false negatives – canonical novels that were not classified as such based on their embeddings. Such an analysis could enhance our understanding of factors such as the role of publishers in the canonization process. Finally, further exploration is needed to understand why certain publishers are so closely linked to canon formation and how their role has evolved over time.

In terms of the methods we used, improving our sampling techniques (both through downsampling and upsampling) and refining our approach to text embeddings could enhance our results. Rather than averaging vectors, alternative approaches could be explored to experiment with different aggregation strategies. Further research is still needed to develop a more comprehensive understanding of what embeddings capture – and what they overlook. This could involve not only comparing embeddings with other textual features, such as syntactic complexity, cognitive processing difficulty, and stylistic patterns, but also employing these features as standalone text-intrinsic measures. Future work could also explore experiments with Generalized Additive Models (GAMs) to analyze potential non-linear relationships between features and classification outcomes, providing a more flexible yet interpretable alternative to linear models. Additionally, simpler and more interpretable methods, such as TF-IDF, could serve both as points of comparison and as alternative ways to analyze textual characteristics.[12]

7 Conclusion

This paper has examined the roles of text-extrinsic and text-intrinsic features in shaping a novel's canonicity, using the Danish Modern Breakthrough era (1870-1900) as a case study. We employed embeddings generated with the multilingual `m-e5-large-instruct` model as text-intrinsic features, while our text-extrinsic features included the novel's price, publisher, and the author's nationality. Using a Random Forest classification model, we predicted whether a novel belonged to the *Canon* or *Other* category based on various feature sets. Our findings demonstrate that text-extrinsic features are strong predictors of a novel's canonicity, suggesting that external dynamics play a significant role in canon formation. At the same time, embeddings alone emerged as robust predictors for canonicity, both on their own and when combined with text-extrinsic features. Importantly, we show that these results are not disproportionately influenced by the many small publishing houses that each published a single non-canonical novel.

We also explored what misclassifications reveal about the boundaries of the literary canon. By focusing on non-canonical novels with textual profiles similar to canonical works, we investigated why these novels failed to achieve canonical status. Our analysis seems to show that, for many authors, text-intrinsic characteristics were insufficient to secure a place in the canon. Conversely, we demonstrated that some novels exhibiting the correct textual profile, publisher, and price still failed to achieve canonical recognition.

Limitations

Creating embeddings

Prompts: This work utilizes the prompt-based embedding model `m-e5-large-instruct`. It is likely that embeddings could be notably different when using a different prompt. The chosen prompt was based on the tests in Feldkamp et al. (2024b), where the prompt '*Identify the author of a given passage from historical Danish fiction*' was used in the clustering task for historical Danish. Further prompt variations and variation effects on embeddings were presented in Feldkamp et al. (2024b).

Occurrence in training data: Canonical works may appear more frequently online or in varied contexts, potentially influencing embeddings in web-trained models. However, this effect is likely minor, as historical novels make up a small fraction of online discourse – especially in Danish, which represents a tiny portion of the multilingual model's training data. Ideally, training data should be examined, but this is often unfeasible due to limited access and computational constraints. The frequent rewriting of historical canons further complicates such efforts.

Canon definition

The concept of canonicity is inherently vague and subject to various interpretations. In our study, we adopt a binary categorization (canon/non-canon) as a pragmatic choice, acknowledging that the boundary between these categories is more fluid than

[12] A comparison between embeddings and TF-IDF representations was included in Feldkamp et al. (2024b).

our classification suggests. Our goal is to estimate broad distinctions rather than capture the full complexity of canon formation.

However, this approach may obscure cases where works occupy an ambiguous position within the literary field or where different actor types exert conflicting influence. In fact, this binary categorization simplifies a phenomenon that may be better represented as a continuous or multi-dimensional variable (Brottrager et al., 2022). One key issue with continuous canon variables is that they often assume independence between different actor evaluations – for instance, treating scholarly recognition and institutional adoption as separate yet equally weighted factors. In practice, these evaluations are often highly collinear, as institutional canons tend to reflect scholarly assessments, and vice versa (Feldkamp et al., 2024a; Barré et al., 2023). A more refined approach would account for these dependencies, potentially assigning different weights based on the extent to which one form of recognition reinforces another.

A further complication is whether canonicity should be treated as a singular phenomenon – one that different actor evaluations, such as scholars, institutions, etc., provide partial windows onto – or as multiple, overlapping but distinct processes. In our case, we implicitly conflate expert and government evaluations, assuming they reflect the same underlying phenomenon of "canon". This may not always hold, and future research could explore whether different forms of recognition should be treated as separate dimensions of canonicity or as interrelated signals of a shared phenomenon.

Acknowledgments

The authors of this paper were supported by grants from the Carlsberg Foundation (*The Golden Array of Danish Cultural Heritage*) and the Aarhus Universitets Forskningsfond (*Golden Imprints of Danish Cultural Heritage*).

Part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

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.

Jean Barré. 2024. Latent Structures of Intertextuality in French Fiction: How literary recognition and subgenres are framing textuality. In *Proceedings of the Computational Humanities Research Conference 2024*, volume 3834, pages 21–36, Aarhus, Denmark. CEUR-WS.

Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. 2023. Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature. *Journal of Cultural Analytics*, 8(3).

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024. Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality. *Preprint*, arXiv:2404.04022.

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.

Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case. *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189.

Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case. *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189. Number: 89.

Jens Bjerring-Hansen and Matthew Wilkens. 2023. Deep distant reading: The rise of realism in Scandinavian literature as a case study. *Orbis Litterarum*, 78(5):335–352.

Harold Bloom. 1995. *The Western Canon: The Books and School of the Ages*, 1st riverhead ed edition. Riverhead Books, New York, NY.

Judith Brottrager, Annina Stahl, and Arda Arslan. 2021. Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features. In *Proceedings of the Computational Humanities Research Conference 2021*, volume 2989, pages 195–205.

Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. Modeling and predicting literary reception. *Journal of Computational Literary Studies*, 1(1):1–27.

Giuliano D'Amico. 2016. Modern Breakthrough. In *Routledge Encyclopedia of Modernism*, 1 edition. Routledge, London.

Kenneth Enevoldsen, Lasse Hansen, Dan S. Nielsen, Rasmus A. F. Egebæk, Søren V. Holm, Martin C. Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B. Jørgensen, Malte Højmark-Bertelsen, Peter B. Vahlstrup, Per Møldrup-Dalum, and Kristoffer Nielbo. 2023. Danish foundation models. *Preprint*, arXiv:2311.07264.

Pascale Feldkamp, Yuri Bizzoni, Mads Rosendahl Thomsen, and Kristoffer L. Nielbo. 2024a. Measuring Literary Quality. Proxies and Perspectives. *Journal of Computational Literary Studies*. Forthcoming.

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.

John Guillory. 1995. *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press.

Steen Harbild, Stefan Hermann, and Steen Lassen, editors. 2004. *Dansk litteraturs kanon: rapport fra kanonudvalget*, 1. udg. ; 1. opl edition. Undervisningsministeriets forlag, København.

Leonhard Herrmann. 2011. System? Kanon? Epoche? Perspektiven und Grenzen eines systemtheoretischen Kanonmodells. In Matthias Beilein, Claudia Stockinger, and Simone Winko, editors, *Kanon, Wertung und Vermittlung*, pages 59–76. DE GRUYTER.

Renate von Heydebrand and Simone Winko. 1996. *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. 1953. Schöningh, Paderborn München.

Harald Ilsøe. 2014. Bogligt undertøj. Lidt om danske bogforsatser ca. 1880-1920. *Fund og forskning i det Kongelige Biblioteks samlinger*, 53:209–209.

Lone Kølle Martinsen. 2012. Bondefrihed og andre verdensbilleder. idehistorisker studier af b.s. ingemanns danmarkshistorie 1824-1836. *Temp - tidsskrift for historie*, 3(5):75–103.

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

Jack Douglas Porter. 2018. Popularity/prestige. Technical report, Stanford Literary Lab.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. Success in books: predicting book sales before publication. *EPJ Data Science*, 8(1):31.

Simone Winko. 2002. Literatur-Kanon als 'invisible hand'-Phänomen. In *Literarische Kanonbildung*, pages 9–24. TEXT+KRITIK.

Yara Wu. 2023. Predicting the Unpredictable. Using Language Models to Assess Literary Quality. Master's thesis, Uppsala University, Uppsala.

Yara 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.

A Number of novels per publishing house

The heatmap with number of novels in each category published by a given publishing house can be found in Figure 3.

B Non-canonical canonical novels

The heatmap with non-canonical novels incorrectly classified as canonical based on text-intrinsic features can be found in Figure 4.

Figure 3: Number of novels in each category published by a given publishing house. Note that overall, the *Other* category generally has a higher entropy in its distribution over publisher than the *Canon* category. Entropy, $Other = 3.66$, $Canon = 1.72$. This difference persists when downsampling the majority group.

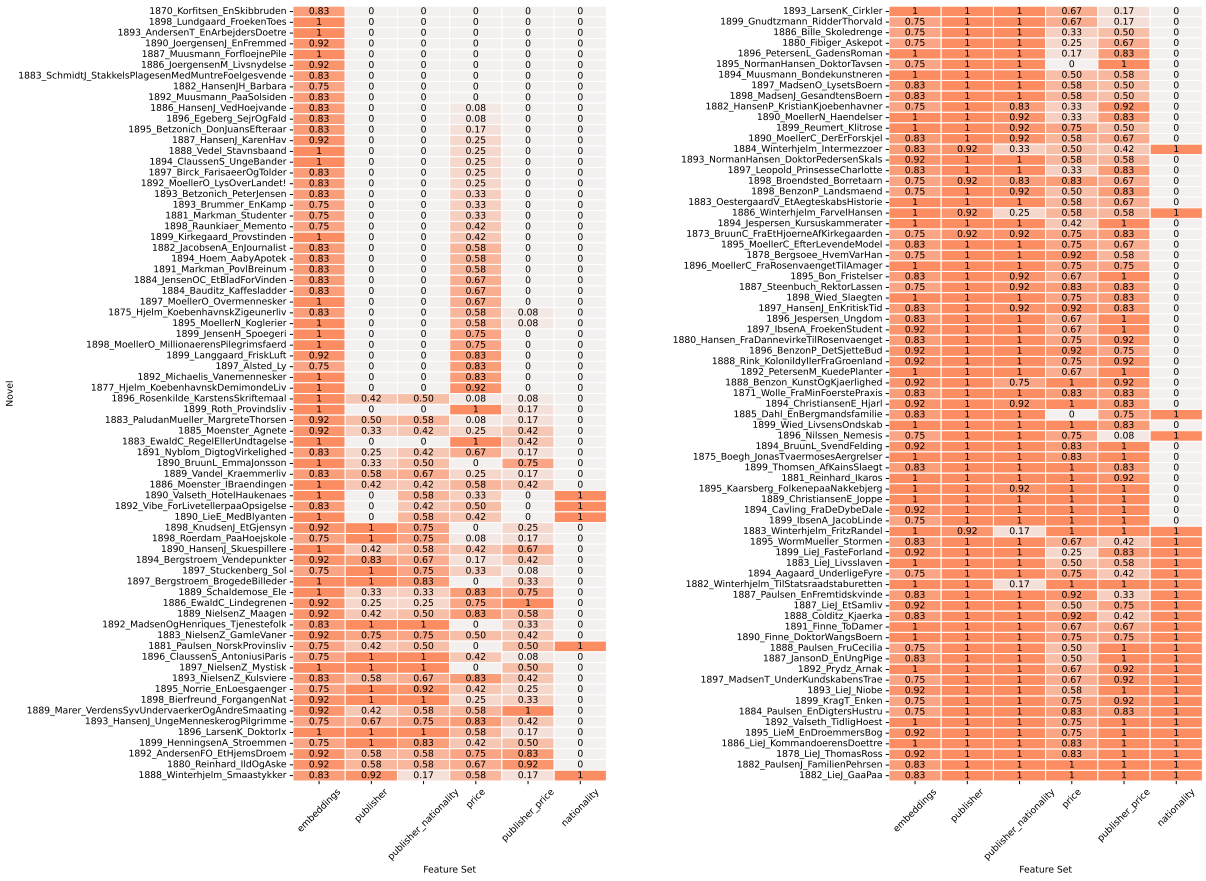

Figure 4: Heatmap of false positives (FP): non-canonical novels incorrectly classified as canonical based on $embeddings$ as features. Columns represent feature sets, with cell values showing normalized false positive counts ($FP/(TN + FP)$. We only include novels of which 75% of the predictions for $embeddings$ are FP ($embeddings >= 0.75$). The unique number of novels is 145, and every novel is predicted 12 times.