Literary Time Travel: Distinguishing Past and Contemporary Worlds in Danish and Norwegian Fiction

Jens Bjerring-Hansen[1], Ali Al-Laith[1,2], Daniel Hershcovich[2], Alexander Conroy[1] and Sebastian Ørtoft Rasmussen[3]

[1]*Department of Nordic Studies and Linguistics, University of Copenhagen*
[2]*Department of Computer Science, University of Copenhagen*
[3]*Department of Comparative Literature and Rhetoric, Aarhus University*

Abstract

The classification of historical and contemporary novels is a nuanced task that has traditionally relied on expert literary analysis. This paper introduces a novel dataset comprising Danish and Norwegian novels from the last 30 years of the 19th century, annotated by literary scholars to distinguish between historical and contemporary works. While this manual classification is time-consuming and subjective, our approach leverages pre-trained language models to streamline and potentially standardize this process. We evaluate their effectiveness in automating this classification by examining their performance on titles and the first few sentences of each novel. After fine-tuning, the models show good performance but fail to fully capture the nuanced understanding exhibited by literary scholars. This research underscores the potential and limitations of NLP in literary genre classification and suggests avenues for further improvement, such as incorporating more sophisticated model architectures or hybrid methods that blend machine learning with expert knowledge. Our findings contribute to the broader field of computational humanities by highlighting the challenges and opportunities in automating literary analysis.

Keywords

Historical Text, Text Classification, Danish, Norwegian, Literature

1. Introduction

Some novels are set in the past, offering readers a glimpse into historical periods, while others reflect the time in which they were written, dealing with contemporary issues. For example, the novel *Lolotte. En Roman fra den Gustavianske Tid* (Lolotte. A Novel from the Gustavian Period, 1898) by Marie Henckel clearly signals its historical nature through its subtitle, which specifies

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

jbh@hum.ku.dk (J. Bjerring-Hansen); alal@di.ku.dk (A. Al-Laith); dh@di.ku.dk (D. Hershcovich); alc@hum.ku.dk (A. Conroy); soer@cc.au.dk (S. Ø. Rasmussen)
https://nors.ku.dk/english/staff/?pure=en/persons/195540 (J. Bjerring-Hansen); https://cst.ku.dk/english/ansatte/?pure=en/persons/741529 (A. Al-Laith); https://danielhers.github.io/ (D. Hershcovich); https://nors.ku.dk/english/staff/?pure=en/persons/195540 (A. Conroy); https://pure.au.dk/portal/da/persons/soer%40cc.au.dk (S. Ø. Rasmussen)
0000-0001-5786-8300 (J. Bjerring-Hansen); 0000-0001-6650-3469 (A. Al-Laith); 0000-0002-3966-8708 (D. Hershcovich); 0000-0001-5786-8300 (A. Conroy); 0000-0002-6238-2513 (S. Ø. Rasmussen)

a late 18th-century setting. In contrast, Albert Gnudtzman's *Ridder Thorvald. En lille københavnsk Roman* (Knight Thorvald. A small Copenhagen novel, 1899) initially misleads with the historical-sounding keyword "knight" in the title, but the opening scene set in a modern urban café distinctly establishes it as a contemporary novel.

The question of whether a novel is set in modern days or historical times (contemporary or historical novel?) was by no means uncontroversial in the time of the so-called Modern Breakthrough in Scandinavian literature circa 1870-1900 [6]. On the contrary, it was a question of taste (good vis-à-vis bad) and, accordingly, a detection and quantification of the historical novel give insight into the cultural divides of the period. Modern realist aesthetics ostracized the historical novel and insisted that literature should be situated in the present and address current problems. In 1871, famously and characteristically, the influential Danish critic Georg Brandes, referring to Scott's Waverley novels from the early 1800s, rejected the historical novel as "an unfortunate and now abandoned genre, imported from Scotland and invented by a pure-blooded Tory, which originated in a state of mind similar to ours, one with all its ideals in the past" [7]. And at least for a while, the historical novel was aesthetically and socially demoted to the realm of popular literature, which no one, except for readers and consumers, cared about. In the 20th century advanced definitions of the historical novel and its complex relationship to its political contexts and to the development of the genre towards realism and modernism have been major points of discussion in the historiography of the novel. In our paper, we approach the question more directly by tentatively putting ourselves in the place of the historical actors, professional tastemakers like Brandes or conventional consumers in the literary marketplace, and emphasizing some immediate, easily decodable genre signals from paratext (titles and subtitles) and text (the opening of the novels).

We introduce a dataset of Danish and Norwegian novels from the last 30 years of the 1800s, annotated by literary scholars according to whether they are historical or contemporary. The novels are taken from the MeMo (Measuring Modernity) corpus [5], comprising 859 novels. We assess the ability of language models to generalize this generic distinction as expressed in their titles and first few sentences, by training them on a portion of the dataset and evaluating them on unseen novels. While fine-tuned Danish language models show good performance in the task, error analysis reveals they still lack sensitivity to salient cues that literary scholars observe.[1]

2. Related Work

Text classification is a pivotal task in natural language processing (NLP) that entails categorizing text into predefined labels or classes. It has a broad spectrum of applications, including sentiment analysis [1], word sense disambiguation [22], named entity recognition [12, 4, 19], and genre classification [32, 21]. With the advent of pre-trained language models like BERT [10], GPT [35], and their variants, significant advancements have been achieved in this domain. These models leverage extensive text corpora to enhance the understanding of context and semantics [27], establishing new standards in accuracy and robustness. Consequently, they enable more nuanced and sophisticated text classification systems capable of handling

[1] Our dataset, code and models will be made publicly available upon publication.

diverse and complex textual data. Current research continues to investigate enhancements in model architectures, fine-tuning techniques, and domain-specific adaptations to further boost the performance of text classification tasks.

When dealing with literature, in academic as well as everyday contexts, taxonomic thinking and practices seem both habitual and inevitable. Literary genre studies got underway with the Ancient Greeks, from which the division of poetic literature in three main genres: lyric, epic and drama, often ascribed to Aristotle and his *Poetics*, has proliferated [17]. Since then an enormous and ever-growing body of genre theory has developed [14]. Of special interest to us is:

1. scholarship on the historical novel of the 19th century which often serves as the predecessor and/or antidote to the modern realist (and contemporary set) novel [25, 13, 40, 47],
2. historical studies, influenced by the sociology of literature and the history of the book, concerned with genre fiction and its aesthetic and commercial development in the 19th century [36, 16], and
3. (non-digital) quantitative approaches to the history of the novel [28, 33, 30, 15].

Within computational literary studies of recent years, genre has been an important touch point for NLP approaches and literary theory and historiography. Text genre classification is a crucial area of research that aids in systematically categorizing vast and diverse collections of literary works. This task involves distinguishing between various genres such as fiction and non-fiction [46, 45, 34], poetry [37], and drama [38], among others, within literary corpora. Also, significant efforts have been made to classify novels in various sub-genres, predominantly with a focus on volume-level similarity across a range of features that capture significant generic aspects [46, 8, 44]. The advancements facilitated by NLP techniques and machine learning, including predictive modeling, are substantial, resulting in more accurate and automated genre classification while also embracing notions from contemporary literary scholarship that a literary genre comprises many features rather than a single defining characteristic [39, 44, 23]. As Ted Underwood has argued, "[t]he best way to measure the differentiation between literary genres is probably to train supervised predictive models that attempt to distinguish works in one genre from other works in a given period or cultural milieu" [39].

3. Historical vs. Contemporary Novels

There is a long and intensive research tradition that has been interested in the political and social-historical implications of the historical novel of the 19th century and the decline of the genre in the latter part of the century as a reflex of a new aesthetic positions with a primacy of immediate perception and contemporaneity [25, 2, 29]. In this context, complex definitions have been drafted on the basis of intensive close readings of particularly British, French and Russian novels. Lukacs' five principal claims about the genre is a pioneering example of this [25]. However, in practice, the question of categorization poses fewer problems for both literary scholars and customers at bookstores. If you consider a common definition of the genre, such as this one from a literary reference work, it will correspond to most readers' common and more or less reflected perception of the genre:

> A novel in which the action takes place during a specific historical period well before the time of writing (often one or two generations before, sometimes several centuries), and in which some attempt is made to depict accurately the customs and mentality of the period.[2]

In this paper, we construct a dataset using the genre classification of the novels of the MeMo corpus performed by Bjerring-Hansen and Ørtoft [6], which has followed such pragmatic and intuitive understanding of the genre as something that can be decoded immediately by a quick inspection of the temporal coordinates of the individual texts. To carry out an analogous quantification of genre trends and proportions between historical and contemporary novels, the authors performed close readings of both (certain) paratexts and (particular) parts of each novel. More specifically, the annotation was carried out on the following premises:

1. Many historical novels "reveal" themselves already in the title (as is the case with *Dronning Caroline Mathilde af Danmark* = Queen Mathilde of Denmark by the pseudonym Caja from 1889) or in the subtitle (as is the case with the anonymously published *Caroline. Bøhmens frygtelige Svøbe eller et Gammel Bjergslots Hemmelighed: Historisk-romantisk Fortælling* = Caroline. Böhmen's terrible scourge or the secret of an old mountain castle; Historical-romantic tale), or more redundantly both the one and the other (cf. H.F. Ewald's *Griffenfeld. Historisk Roman* = Griffenfeld. Historical Novel from 1888).
2. If the titles do not contain clear paratextual signals, genre affiliation is often indicated on the first few pages of the novel (as, for example, in the case of *Indianerpigen fra Cape Breton* = The Indian Girl from Cape Breton by "L.M", which, although the title page does not indicate that we are dealing with a historical novel, immediately sets the temporal scene with the opening sentence: "It was in the year 1780 [...]").

So, generally, it is striking to what extent the historical novels of the 19th century clearly and actively give away their generic affiliation. As several literary scholars have pointed out, this is probably because the historical novel's foremost characteristic—and selling point—is its historical setting, which, then, producers and distributors clearly wants to mark for the intended readership and therefore already on the title page or the first pages "come clean" [47, 42]. It can be added that these guidelines only to a very limited extent can be reversed on the basis of a similar "scanning" of the paratextual and textual evidence. Non-historical, i.e. contemporary novels—the novels which aesthetics and the criticism in the late 1800s placed a decisive and favourable emphasis on—do not communicate their temporality in a similar way. The MeMo corpus entails a few handfuls of instances of emphatically contemporary subtitles (e.g. "Nutidsfortælling" = story from the present day, "Samtidsroman" = contemporary novel etc.), but in general the contemporary novels are implying their genericity through silence on their temporal setting.

This transparent literary communication, or consumer information, which is of course less obvious in the few and often canonized instances of experimental novels that "play" with genre fiction such as the historical novel, can be said to be a general feature of popular literature, including also romances and detective stories etc., and the genre-fiction system, developing

[2] https://www.oxfordreference.com/display/10.1093/oi/authority.20111104173823536

Table 1
MeMo corpus statistics.

Total novels	859
Total sentences	3,282,643
Total words	53,588,381
Average sentences per novel	3,821
Average words per novel	62,385
Average words per sentence	16.3

in the latter part of the 19th century [28, 16]. The question is whether machines can learn to read these literary and cultural signals, apparent in the paratext and/or the opening pages of the novel, which for historical actors have seemed quite obvious? (Non-)Historical novel – yes or no? In other words, the genre distinctions that our method rely on are historically framed, meaning they are tied to specific periods and cultural contexts rather than having universal relevance across time and place.

4. Methodology

To address this question, we treat the problem from a machine learning perspective. We introduce an annotated corpus and fine-tune pre-trained transformer language models on it, evaluating their performance on a held-out test set.

4.1. Dataset

We rely on the MeMo corpus [5], comprising 859 Danish and Norwegian novels spanning the last 30 years of the 19th century, with more than 64 million tokens. The corpus is a rich and diverse collection of texts that provides valuable insights into the classification of novels as historical and contemporary during the period under investigation. Table 1 shows statistical information about the corpus. We obtain the annotated dataset of novels from Bjerring-Hansen and Ørtoft [6]. The final list of annotated novels consists of 859 novels, with 78% categorized as contemporary and 22% as historical. Figure 1 illustrates the temporal distribution of historical and contemporary novels in our corpus.

4.2. Novel Classification

We use the dataset for training and evaluating transformer-based language models. Specifically, inspired by the observations made by Bjerring-Hansen and Ørtoft, we consider three settings:

1. Providing the title and sub-title of the novel as input to the model,
2. Providing the first 15 sentences of the novel as input to the model,
3. Concatenating the title, sub-title and first 15 sentences and providing them to the model.

In all cases, we train the model to classify the novel according to the binary label obtained from the annotated dataset. Subsequently, we evaluate the models on a test set of held-out novels to assess their ability to generalize the ability to identify the cues learned during training.

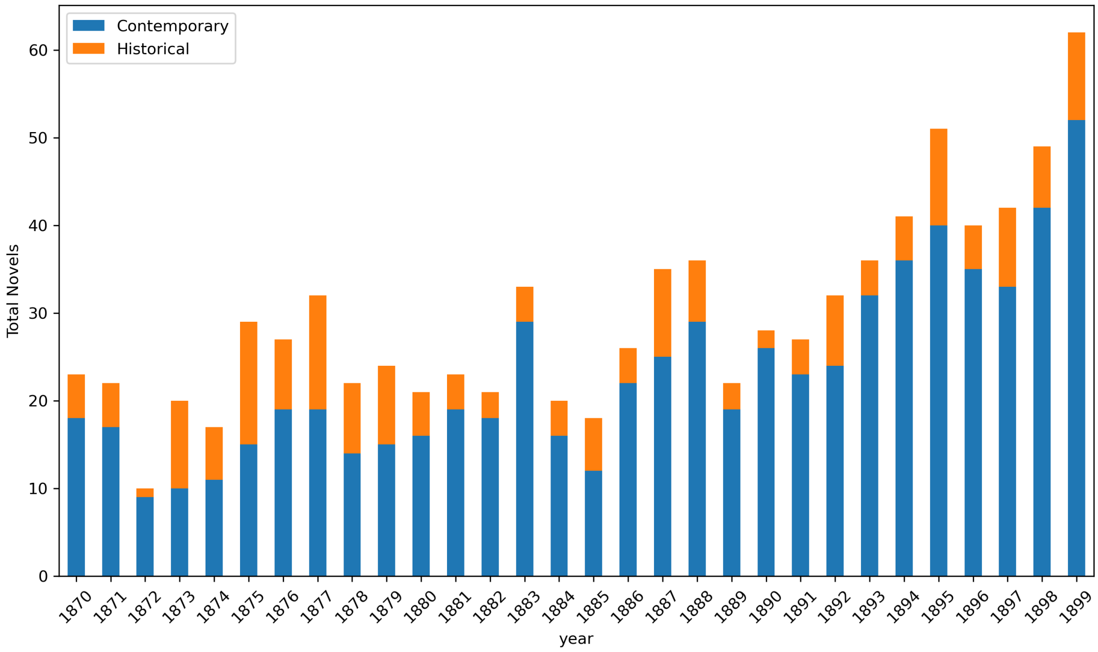

Figure 1: Distribution of Historical and Contemporary Novels in the MeMo Corpus Over Time.

5. Experiments and Results

We experiment with four pre-trained language models and three types of provided context, comparing their performance and using them as the basis for an elaborate analysis of errors and indicative features.

5.1. Pre-trained Language Models

The models evaluated in our novel classification experiments had been pre-trained on text corpora including Danish and Norwegian text. We train them on the task using supervised fine-tuning. Importantly, all models are selected based on their performance evaluated on Danish and Norwegian literary benchmark datasets [22], the Scandinavian Embedding Benchmark[3] and ScandEval,[4] [31] even though these models had not been trained primarily on historical Danish or Norwegian. We additionally experiment with a model (MeMo-BERT-03) specifically adapted for the MeMo corpus.

DanskBERT. DanskBERT,[5] a top-performing Danish language model noted for its success on the ScandEval benchmark [41], is based on the XLM-RoBERTa architecture and trained on the Danish Gigaword Corpus [43]. It features 24 layers, a hidden dimension of 1024, 16

[3] https://kennethenevoldsen.github.io/scandinavian-embedding-benchmark/
[4] https://scandeval.com/
[5] https://huggingface.co/vesteinn/DanskBERT

attention heads, and a subword vocabulary of 250,000. The model was trained with a batch size of 2,000 for 500,000 steps on 16 V100 GPUs over two weeks.

Danish Foundation Models sentence encoder. A sentence-transformers model [11] based on the BERT architecture, featuring 24 layers, 16 attention heads, and a hidden size of 1024. It incorporates a dropout rate of 0.1 for attention probabilities and hidden states, using GELU activation and supporting up to 512 position embeddings. With a vocabulary size of 50,000 tokens, this model, referred to as DFM (Large), excels in tasks such as Danish sentiment analysis and named entity recognition.[6]

MeMo-BERT-03. Developed by continuing the pre-training of the pre-trained Transformer language model DanskBERT [22].[7] This foundation allows MeMo-BERT-3 to leverage extensive linguistic knowledge for NLP tasks in historical literary Danish including sentiment analysis and word sense disambiguation. The model outperformed different models in sentiment analysis and word sense disambiguation tasks [22].

NB-BERT-base. A general-purpose BERT-base model was developed using the extensive digital collection at the National Library of Norway [20].[8] It follows the architecture of the BERT Cased multilingual model and has been trained on a diverse range of Norwegian texts, encompassing both Bokmål and Nynorsk from the past 200 years. This comprehensive training allows the NB-BERT-base to effectively handle a wide array of NLP tasks in Norwegian. The model achieved the second-highest performance ranking in the Norwegian Named Entity Recognition task compared to other models listed on the ScandEval benchmark for Norwegian natural language understanding.

5.2. Experimental Setup

Our experiments involve fine-tuning the pre-trained language models on the annotated novels from our corpus. To enable testing of generalization in the face of temporal shift [26], the last 130 novels according to publication year (≈15%) are used as a testing set, while the remaining novels were randomly divided into training and validation with 70% and 15% respectively. The experiments involve fine-tuning the models on the dataset using a batch size of 32, training for 20 epochs with the AdamW optimizer [24] at a learning rate of 10^{-3}. During training, we monitor the performance on the validation set to assess model convergence and to prevent overfitting, keeping the checkpoint with the best validation score. For evaluation, we employ the F1-score metric due to its ability to balance precision and recall, particularly effective for tasks with imbalanced datasets. The performance of each model is evaluated on both validation and test sets, ensuring the robustness and generalizability of the models across different datasets and epochs. For comparison, due to the imbalanced nature of the dataset, with 22% of novels being historical overall and the percentage being 17% in the test set, a naive baseline that selects a label based on the training distribution would achieve about 70% weighted F1-score.

[6] https://huggingface.co/KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align

[7] https://huggingface.co/MiMe-MeMo/MeMo-BERT-03

[8] https://huggingface.co/NbAiLab/nb-bert-base

Table 2
Fine-tuning Classification Results: F1-score for the four pre-trained Transformer language models in three input settings on both validation and tests sets.

	Titles & Sub-titles		First 15 Sentences		Both	
Model	**Valid.**	**Test**	**Valid.**	**Test**	**Valid.**	**Test**
DanskBERT	0.91	0.88	0.80	0.82	0.81	**0.84**
DFM (Large)	**0.92**	0.88	**0.81**	**0.86**	**0.82**	0.83
MeMo-BERT-03	0.89	**0.91**	**0.81**	0.85	**0.82**	0.83
NB-BERT-base	0.91	0.89	0.79	0.83	**0.82**	**0.84**

5.3. Novels Classification Experiments

5.3.1. Titles and Sub-titles Classification

In this experiment, we concatenate the title and subtitle of each novel and perform classification by fine-tuning the aforementioned pre-trained language models with the novel labels, using the cross-entropy objective. Table 2 (left) presents the fine-tuning results of the selected models. DFM (Large) achieved the highest performance on the validation set with an F1-score of 92%, while the MeMo-BERT-03 model excelled on the testing set with an F1-score of 91%.

5.3.2. First 15 Sentences Classification

We use the Danish pipeline in spaCy [18] for sentence segmentation and extract the first 15 sentences from each novel. We then use each sentence as a separate input instance for fine-tuning the aforementioned pre-trained language models, with the same novel-level labels as previously now inducing sentence-level labels. To predict novel-level labels using the fine-tuned models, we apply them to the first 15 sentences of a (validation or testing) novel, and use majority voting to determine the novel-level predictions.

The results of fine-tuning the models is shown in Table 2 (middle). DFM (Large) and MeMo-BERT-03 achieved the highest performance on the validation set with an F1-score of 81%, while DFM (Large) excelled on the testing set with an F1-score of 86%. Notably, for all models, using the first 15 sentences as input performs worse than using the title and sub-title.

5.3.3. Both Titles & Sub-titles and First 15 Sentences Classification

In this experiment, we combine both the titles & sub-titles and the first 15 sentences of each novel in the corpus: technically, we repeat the same setup as using the first 15 sentences, but additionally prepend the concatenated title and sub-title as if they were another sentence. The fine-tuning results of the four models of this experiments are shown in Table 2 (right). DFM (Large), MeMo-BERT-03 and NB-BERT-Base achieve equal performance on the validation set with an F1-score of 82%, while DanskBERT and NB-BERT-Base perform best on the testing set with an F1-score of 84%. Overall, performance in this setting is similar to just using the first 15 sentences, but the best performance on the test set is in fact obtained when just using titles and sub-titles, and ignoring the first 15 sentences.

Table 3
Expected Calibration Error (ECE) for the models on the test set.

Model	Titles & Sub-titles	First 15 Sentences	Both
DFM (Large)	0.040	0.085	0.094
DanskBERT	0.043	0.087	0.081
MeMo-BERT-03	0.062	**0.076**	0.098
NB-BERT-base	**0.028**	0.156	**0.070**

6. Discussion

While all models are highly accurate after fine-tuning, surprisingly, we observe that the best predictions are obtained by just using the title and sub-title as input, disregarding the first 15 sentences of the novel. This suggests either that the genre information is less salient in the first 15 sentences, or that the models are not as capable of extracting it from them. To analyze this further, we investigate model confidence on mislabeled predictions, and perform a fine-grained error analysis.

6.1. Model Calibration

When reading the text opening, introspection from expert annotation reveals that genre identification often hinges on specific key sentences (e.g., mentioning specific entities, events, or years) rather than the entire opening passage. While most sentences in the opening text do not clearly suggest one genre or another, these ”giveaways” are sparse but salient. This nuance may be lost by the majority voting procedure over sentences, leading to misclassifications when the first 15 sentences are used as input. Therefore, we are interested in the model's confidence and whether it is calibrated to match the experts' uncertainty or disagreement about the labels [3].

To evaluate model calibration, we use Expected Calibration Error (ECE), a metric that measures how well the model's predicted probabilities reflect the true accuracy [9]:

$$ECE = \sum_{k=1}^{K} \frac{|B_k|}{n} |acc(B_k) - conf(B_k)|$$

where $K = 10$ is the number of bins (confidence intervals), $|B_k|$ is the number of samples in bin k, $acc(B_k)$ is the accuracy in bin k, and $conf(B_k)$ is the average confidence in bin k.

When using titles and sub-titles as input, the best-performing model (MeMo-BERT-03) achieved a relatively low ECE of 0.062, as shown in Table 3, indicating that it was reasonably well-calibrated when relying on paratextual information. Titles and sub-titles often contain clear genre markers that allow the model to make high-confidence, mostly accurate predictions. However, the model still made misclassifications, particularly when historical-sounding titles misled the model. Despite this, the model's overall confidence generally matched its performance in this setting, and it exhibited the lowest calibration error compared to other settings. This result underscores the strength of using paratextual clues, though it also reveals that misleading terms in the title can cause overconfidence in wrong predictions.

In contrast, when the models used the first 15 sentences of the novels as input, calibration worsened across all models. For example, the ECE for DFM (Large) increased to 0.085, and other models similarly struggled with higher calibration errors (see Table 3). This is likely because, as noted in expert analyses, genre signals are not uniformly distributed across the opening sentences. Instead, they tend to appear in specific key sentences that reveal important genre-relevant details. In many cases, the first 15 sentences are ambiguous, lacking explicit time markers or character descriptions, which increases the model's uncertainty. However, rather than reflecting this uncertainty in their confidence scores, the models often exhibited overconfidence, resulting in higher calibration errors. This overconfidence indicates that the models are not adequately capturing the uncertainty present in the text openings, a discrepancy that reflects the challenge of extracting nuanced genre information from longer inputs.

Combining both titles, sub-titles, and the first 15 sentences of the novels did not uniformly improve calibration. For MeMo-BERT-03, the ECE increased to 0.094, suggesting that integrating both sources of information did not lead to better confidence alignment. Although the models had access to more context, they struggled to effectively weigh the paratext against the more ambiguous textual cues from the opening sentences. In some cases, conflicting signals between the title and the text may have caused the models to oscillate between genres, ultimately leading to poorer calibration. DanskBERT, however, exhibited slightly better calibration in this setting, indicating that it was more adept at integrating the two types of input compared to other models. This slight improvement over single-input settings suggests that certain models can benefit from additional context, though the integration process remains challenging for most.

6.2. Error Analysis

We discuss the errors encountered during the classification of historical and contemporary novels, focusing on prediction errors made by the best-performing models, as well as annotation errors identified by expert inspection of the mislabeled predictions.

6.2.1. Prediction Errors

A notable pattern in the prediction errors is the tendency of the models to misclassify contemporary novels as historical based on titles (in 12 out of 17 cases where at least one model misclassified the label in this setting) and based on text openings (first 15 sentences) or the combination of titles and text openings (11 out of 11 cases). An illustrative example of the first type of error is Albert Gnudtzman's urban novel *Ridder Thorvald. En lille københavnsk Roman* (Knight Thorvald. A small Copenhagen novel, 1899). The title's keyword "knight" leads the models to misinterpret it as a historical romance. However, the opening scene set in a lively urban café correctly classifies it as a contemporary novel, highlighting the discrepancy between title-based and content-based classification.

When models misinterpret text openings (first 15 sentences), a common issue is their failure to recognize historical settings established through character introductions rather than explicit time clues. For instance, in Marie Henckel's *Lolotte. En Roman fra den Gustavianske Tid* (Lolotte. A Novel from the Gustavian Period, 1898), while the subtitle clearly indicates a late 18th-century setting, the models fail to date the characters like Prince Gustaf and Sofie Magdalene, leading to incorrect classifications.

Machine readings of genre clues also shed light on borderline cases, such as novels set in the near past relative to the modern breakthrough period (1870-99). Examples include novels set during the Danish-German wars of 1848-50 and 1864, like Chr. Christensen's *Kærlighedens Mysterier. En Historie fra 1848-50* (The Mysteries of Love. A Tale from 1848-50, 1899) and P.A. Worm's *Forbrydelsernes Konge eller Den skalperede Præst* (The King of Crime and the Scalped Priest, 1899). An intriguing case involves a novel beginning in the narrator's present with modern elements like electric light and a telephone but transitioning to a historical analepsis: U. Ravn's *Interioerer fra vores Bedsteforældres Tid* (Interiors from the Time of our Grandparents, 1899).

6.2.2. Annotation Errors

After in-depth expert analysis, eight of the models' "errors" in the test set turned out to be Bjerring-Hansen and Ørtoft's annotation errors, including four plain mistakes and four tricky in-between novels where further inspection validated the models' predictions. These erroneous annotations were evenly distributed between misclassified historical and contemporary novels.

An interesting case is the novel *Hvorfor hun blev Nonne. En Fortælling om fransk Klosterliv* (Why she became a nun. A story about French monastic life, 1899) by the pseudonym "Herdis". Both models and annotators were misled by the title into thinking it was a medieval story. However, a close reading of the text's opening pages revealed it to be set in modern times, aligning with the neo-romantic current of the 1890s that revived the historical novel and Catholic themes.

7. Conclusion

In this study, we presented a dataset of Danish and Norwegian novels from the late 19th century, classified as historical or contemporary by literary scholars. We investigated the performance of several pre-trained language models in distinguishing between these two genres based on titles and the first few sentences. While the models demonstrated commendable accuracy, the error analysis revealed limitations in capturing the nuanced cues recognized by human experts. These findings underscore the complexity of literary text classification and suggest that while NLP models can significantly aid in the categorization process, they still require further refinement to match human interpretative abilities fully. In our approach, we chose to limit the textual input to only the titles and the first 15 sentences of each novel. This decision was informed by the pretraining of the models, which predominantly focused on non-literary and contemporary content, meaning that historical figures, settings, and subtleties were likely underrepresented in the training data. As a result, we hypothesized that the opening framing of the novels, which often serves to establish the genre, would be more effective for detection than deeper content. This was evident in our findings, where the titles and subtitles emerged as the most straightforward indicators of the genre distinction. While expanding the analysis to include more content from the novels could potentially capture stronger genre signals, the results indicated that the models performed best on the titles and subtitles; in fact, the first 15 sentences did not surpass the effectiveness of the titles alone. This suggests that the model, like the historical readers, effectively identifies key genre signals from surface-level clues, such as titles and subtitles. While our method is aligned with historical signals by focusing on initial clues that reflect the genre distinctions recognized during the period, it also highlights the need for deeper content analysis, which would require models pretrained on a substantial amount of historical material extending well before the end of the 19th century. To address this, future work could explore more sophisticated model architectures or hybrid approaches that combine machine learning with expert knowledge to enhance the accuracy and depth of genre classification in literary studies.

Our study shows that the historical novel was by no means an extinct genre in Danish-Norwegian literature at the end of the 19th century, as the prevailing modern aesthetic would have it, and furthermore that this was a rather obvious fact, since genre decoding–historical novel – yes or no? – is a relatively trivial affair. It can be determined with great certainty, both by people and models, by reading the paratext and/or the first lines of the novels. Of course, this generic stability cannot be taken for granted or universalized if the fine-tuned models from our study are applied to other textual sources, such as 20th century novels, where genre innovations are increasing and where the historical novel is also exposed to modernist experiments (an early Danish example of this is Nobel Prize winner Johannes V. Jensen's novel *Kongens Fald* (The fall of the king, 1900-01), which represents both a historical depiction of the dramatic events leading to the fall of the Kalmar Union and a modern *flâneur* novel. In future literary studies, we will be able to test the stability of the genre distinctions created during the 19th century, when the literary field and the formation of taste were established, but for accuracy in prediction, we will most likely have to adjust our methods to consider the content of novels on a broader scale.

The comparison between qualitative annotations and machine predictions enhanced our understanding of the quantitative arguments applicable to the period's literature. It highlighted the coexistence and interaction of old and new forms and meanings within an aesthetic timespan. This approach aligns with a broader perspective on genre classification that leverages predictive modeling. As Ted Underwood suggests, our objective shifts from defining a genre to developing a model that can replicate the judgments made by specific historical observers [44]. This paradigm not only advances our technical capabilities but also deepens our literary and historical understanding, bridging the gap between computational methods and humanistic inquiry.

References

[1] A. Allaith, K. Degn, A. Conroy, B. Pedersen, J. Bjerring-Hansen, and D. Hershcovich. "Sentiment Classification of Historical Danish and Norwegian Literary Texts". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Ed. by

T. Alumäe and M. Fishel. Tórshavn, Faroe Islands: University of Tartu Library, 2023, pp. 324–334. URL: https://aclanthology.org/2023.nodalida-1.34.

[2] P. Anderson. "From progress to catastrophe". In: *London Review of Books* 33.15 (2011), pp. 24–28.

[3] J. Baan, W. Aziz, B. Plank, and R. Fernandez. "Stop Measuring Calibration When Humans Disagree". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 1892–1915. DOI: 10.18653/v1/2022.emnlp-main.124. URL: https://aclanthology.org/2022.emnlp-main.124.

[4] D. Bamman, S. Popat, and S. Shen. "An annotated dataset of literary entities". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 2138–2144.

[5] J. Bjerring-Hansen, R. D. Kristensen-McLachlan, P. Diderichsen, and D. H. Hansen. "Mending Fractured Texts. A heuristic procedure for correcting OCR data". In: (2022).

[6] J. Bjerring-Hansen and S. Ø. Rasmussen. "Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case". In: *Passage-Tidsskrift for litteratur og kritik* 38.89 (2023), pp. 171–189.

[7] G. Brandes and L. R. Wilkinson. "The 1872 Introduction to Hovedstrømninger i det 19de Aarhundredes Litteratur (Main Currents of Nineteenth-Century Literature)". In: *PMLA/Publications of the Modern Language Association of America* 132.3 (2017), pp. 696–705. DOI: 10.1632/pmla.2017.132.3.696.

[8] J. Calvo Tello. *The novel in the Spanish Silver Age: a digital analysis of genre using machine learning*. Bielefeld University Press, 2021.

[9] S. Desai and G. Durrett. "Calibration of Pre-trained Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, 2020, pp. 295–302. DOI: 10.18653/v1/2020.emnlp-main.21. URL: https://aclanthology.org/2020.emnlp-main.21.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[11] K. Enevoldsen, L. Hansen, D. S. Nielsen, R. A. F. Egebæk, S. V. Holm, M. C. Nielsen, M. Bernstorff, R. Larsen, P. B. Jørgensen, M. Højmark-Bertelsen, P. B. Vahlstrup, P. Møldrup-Dalum, and K. Nielbo. *Danish Foundation Models*. 2023. arXiv: 2311.07264 [id='cs.CL' full_name='Computation and Language' is_active=True alt_name='cmp-lg' in_archive='cs' is_general=False description='Covers natural language processing. Roughly includes material in ACM Subject Class I.2.7. Note

that work on artificial languages (programming languages, logics, formal systems) that does not explicitly address natural-language issues broadly construed (natural-language processing, computational linguistics, speech, text retrieval, etc.) is not appropriate for this area.'].

[12] A. Erdmann, C. Brown, B. Joseph, M. Janse, P. Ajaka, M. Elsner, and M.-C. de Marneffe. "Challenges and solutions for Latin named entity recognition". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH).* 2016, pp. 85–93.

[13] A. Fleishman. *The English historical novel.* Johns Hopkins University Press, 1971.

[14] J. Frow. *Genre.* Routledge, 2014.

[15] G. Furuland. "Romanen som vardagsvara: förläggare, författare och skönlitterära häftesserier i Sverige 1833-1851 från Lars Johan Hierta till Albert Bonnier". PhD thesis. 2007.

[16] A. Goldstone. "Origins of the US genre-fiction system, 1890–1956". In: *Book history* 26.1 (2023), pp. 203–233.

[17] S. Halliwell. *Aristotle's poetics.* University of Chicago Press, 1998.

[18] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. "spaCy: Industrial-strength Natural Language Processing in Python". In: (2020). DOI: 10.5281/zenodo.1212303.

[19] E. Kogkitsidou and P. Gambette. "Normalisation of 16th and 17th century texts in French and geographical named entity recognition". In: *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities.* 2020, pp. 28–34.

[20] P. E. Kummervold, J. De la Rosa, F. Wetjen, and S. A. Brygfjeld. "Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa).* Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 20–29. URL: https://aclanthology.org/2021.nodalida-main.3.

[21] D. Kurbanova. "Genre Classification and the Current State of Turkmen Musical Folklore". In: *Culture and Arts in the Modern World* 24 (2023), pp. 155–167.

[22] A. Al-Laith, A. Conroy, J. Bjerring-Hansen, and D. Hershcovich. "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).* Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue. Torino, Italia: ELRA and ICCL, 2024, pp. 4811–4819. URL: https://aclanthology.org/2024.lrec-main.431.

[23] S. Liu, Z. Huang, Y. Li, Z. Sun, J. Wu, and H. Zhang. "DeepGenre: Deep Neural Networks for Genre Classification in Literary Works". 2024.

[24] I. Loshchilov and F. Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations.* 2017. URL: https://api.semanticscholar.org/CorpusID:53592270.

[25] G. Lukács. "Der Historische Roman. 1937". In: *Berlin: Aufbau-Verlag* (1955).

[26] J. Lukes and A. Søgaard. "Sentiment analysis under temporal shift". In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* Ed. by A. Balahur, S. M. Mohammad, V. Hoste, and R. Klinger. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 65–71. DOI: 10.18653/v1/W18-6210. URL: https://aclanthology.org/W18-6210.

[27] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. "Recent advances in natural language processing via large pre-trained language models: A survey". In: *ACM Computing Surveys* 56.2 (2023), pp. 1–40.

[28] F. Moretti. "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740?1850)". In: *Critical Inquiry* 36.1 (2009), pp. 134–158. DOI: 10.1086/606125.

[29] R. Mucignat. "Fredric Jameson. The Antinomies of Realism. London: Verso, 2013, 326 pp." In: *Orbis Litterarum* 71.5 (2016), pp. 430–431.

[30] E. Munch-Petersen. "Romanens århundrede: studier i den masselæste oversatte roman i Danmark 1800-1870". In: *(No Title)* (1978).

[31] D. S. Nielsen. "Scandeval: A benchmark for Scandinavian natural language processing". In: *arXiv preprint arXiv:2304.00906* (2023).

[32] J. A. Nolazco-Flores, A. V. Guerrero-Galván, C. Del-Valle-Soto, and L. P. Garcia-Perera. "Genre Classification of Books on Spanish". In: *IEEE Access* 11 (2023), pp. 132878–132892.

[33] N. D. Paige. *Technologies of the Novel: Quantitative Data and the Evolution of Literary Systems.* Cambridge University Press, 2020.

[34] M. R. Qureshi, S. Ranjan, R. Rajkumar, and K. Shah. "A simple approach to classify fictional and non-fictional genres". In: *Proceedings of the Second Workshop on Storytelling.* 2019, pp. 81–89.

[35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language Models are Unsupervised Multitask Learners". In: (2019).

[36] J. A. Radway. *Reading the romance: Women, patriarchy, and popular literature.* Univ of North Carolina Press, 2009.

[37] G. Rakshit, A. Ghosh, P. Bhattacharyya, and G. Haffari. "Automated analysis of Bangla poetry for classification and poet identification". In: *Proceedings of the 12th international conference on natural language processing.* 2015, pp. 247–253.

[38] A. Schneider and P. R. Fabo. "Stage Direction Classification in French Theater: Transfer Learning Experiments". In: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024).* 2024, pp. 278–286.

[39] A. Sharmaa, Y. Hu, P. Wu, W. Shang, S. Singhal, and T. Underwood. "The rise and fall of genre differentiation in English-language fiction". In: *DH2020 (ADHO) Proceedings* 1613 (2020), p. 0073.

[40] H. E. Shaw. *The forms of historical fiction: Sir Walter Scott and his successors*. Cornell University Press, 1983.

[41] V. Snæbjarnarson, A. Simonsen, G. Glavaš, and I. Vulić. "Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: Linköping University Electronic Press, Sweden, 2023.

[42] L. Søndergaard. "At fortælle historier om historien: Om den historiske roman i relation til Poul Vads Rubruk (1972) og Ib Michaels Troubadurens lærling (1983)". In: *Fortællingen i Norden efter 1960*. Aalborg Universitetsforlag, 2004, pp. 404–412.

[43] L. Strømberg-Derczynski, M. Ciosici, R. Baglini, M. H. Christiansen, J. A. Dalsgaard, R. Fusaroli, P. J. Henrichsen, R. Hvingelby, A. Kirkedal, A. S. Kjeldsen, C. Ladefoged, F. Å. Nielsen, J. Madsen, M. L. Petersen, J. H. Rystrøm, and D. Varab. "The Danish Gigaword Corpus". In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 413–421. URL: https://aclanthology.org/2021.nodalida-main.46.

[44] T. Underwood. "The life cycles of genres". In: (2016).

[45] T. Underwood, D. Bamman, and S. Lee. "The transformation of gender in English-language fiction". In: (2018).

[46] M. Wilkens. "Genre, computation, and the varieties of twentieth-century US fiction". In: *Journal of Cultural Analytics* 2.2 (2016).

[47] M. Winge. *Fortiden som spejl*. Lindhardt og Ringhof, 2016.