Noise, Novels, Numbers. A Framework for Detecting and Categorizing Noise in Danish and Norwegian Literature

Ali Al-Laith[1,2], **Daniel Hershcovich**[2], **Jens Bjerring-Hansen**[1],
Jakob Ingemann Parby[3],**Alexander Conroy**[1] , **Timothy R Tangherlini**[4]
Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark[1]
Department of Computer Science, University of Copenhagen, Denmark[2]
Copenhagen Museum, Denmark[3], University of California, Berkeley, USA[4]
alal@di.ku.dk, dh@di.ku.dk, jbh@hum.ku.dk
jakobp@kk.dk, alc@hum.ku.dk, tango@berkeley.edu

Abstract

We present a framework for detecting and categorizing noise in literary texts, demonstrated through its application to Danish and Norwegian literature from the late 19th century. Noise, understood as "aberrant sonic behaviour," is not only an auditory phenomenon but also a cultural construct tied to the processes of civilization and urbanization. We begin by utilizing topic modeling techniques to identify noise-related documents, followed by fine-tuning BERT-based language models trained on Danish and Norwegian texts to analyze a corpus of over 800 novels. We identify and track the prevalence of noise in these texts, offering insights into the literary perceptions of noise during the Scandinavian "Modern Breakthrough" period (1870-1899). Our contributions include the development of a comprehensive dataset annotated for noise-related segments and their categorization into human-made, non-human-made, and musical noises. This study illustrates the framework's potential for enhancing the understanding of the relationship between noise and its literary representations, providing a deeper appreciation of the auditory elements in literary works, including as sources for cultural history.

1 Introduction

Noise, understood as "deviant sonic behaviour" (Novak and Sakakeeny, 2015), is an auditory phenomenon, but also a cultural construct, closely tied to processes of civilization and urbanization. The representation of noise in literature provides insight into the social and cultural developments of the period(s) in which that literature was written.

But can we trace how the past may have sounded? In studies of literature and sound, the empirical foundation is often a small selection of texts, representing either canonical and/or avant-garde instances of 20th century modernism (e.g., Toth, 2016; Lambrecht, 2017; Frattarola, 2018). In our investigations of the soundscapes in the literature of the Scandinavian "Modern Breakthrough" (1870-1899), we broaden the empirical and cultural scope by reading at scale. For that, we develop a framework for the detection and categorization of noise in literary texts.

We extract a series of noise-related topics from a corpus containing more than 800 Danish and Norwegian novels. By examining changes in the frequency of these topics, we draw insight into how literary perceptions of noise have evolved. The findings of this study contribute to a deeper understanding of the relationship between noise and its various representations in literary contexts, and specifically, how 'loud' the literary past was.

Contributions. Our contributions are: (1) the development of a robust and scalable framework for detecting and categorizing noise in literary texts using fine-tuned language models and topic models; (2) the creation of an annotated dataset derived from over 800 Danish and Norwegian novels from the late 19th century, with detailed classifications of noise-related segments into human-made, non-human made, and musical noises; (3) the implementation of analysis to track the evolution of noise perceptions in literature over time, reflecting the cultural and social changes in the soundscapes of the Scandinavian 'Modern Breakthrough'; and (4) a demonstration of the framework's applicability and effectiveness in literary studies, paving the way for future research in other linguistic and cultural contexts. The code and datasets are available under this link: `https://github.com/mime-memo/Noise`.

2 Related Work

Topic modeling on diachronic text. Various methods have been explored for topic modeling in digital humanities and computational literary analysis, specifically in diachronic analysis. Challenges in analyzing diachronic data using topic

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3344–3354
November 12-16, 2024 ©2024 Association for Computational Linguistics

models are discussed in Marjanen et al. (2020), which presents a method for applying topic models to large and imbalanced collections. Heyer et al. (2017) explore detecting semantic change and term evolution through three approaches and introduce "context volatility" as a measure for detecting semantic change. The "Draw My Topics" toolkit, presented by Dou et al. (2016), uses an algorithm based on Vector Space Model and Conditional Entropy to incorporate social scientists' interests into standard topic modeling. Sirin and Lippincott (2024) presented a unique mix of dynamic embedded topic models and change-point detection to analyze the temporal shifts in lexical semantic modality in classical and early Christian Latin texts, offering methods for pattern analysis and integration with traditional literary scholarship, while suggesting future improvements for adapting to less curated materials. A dynamic embedded topic model (D-ETM) combines dynamic latent Dirichlet allocation and word embeddings is introduced by Dieng et al. (2019) to capture evolving word patterns over time, demonstrating superior performance in document completion tasks and topic coherence across diverse datasets, with improved efficiency in model fitting. Haider (2019) use Latent Dirichlet Allocation (LDA) for distant reading tasks on literary data, such as classifying poems by time periods and authorship attribution, while Tangherlini and Leonard (2013) show how LDA can be used as part of a search and discovery pipeline for literary study and the emergence of topics across domains. Finally, Roberts et al. (2013)'s Structural Topic Models use metadata to generate topic prevalence and have had a substantial impact in the social sciences.

Noise/sound in literary texts. A strand of inquiry close to our empirical approach focuses on the representation of sounds and their associated soundscapes within historical and fictional worlds. Schafer (1993) pioneers this approach by examining evolving soundscapes, both real-world and fictional, emphasizing the role of writers as "earwitnesses" to past sonic environments. Recent studies extend this exploration to fictional soundscapes, exploring how descriptions of ambient sounds contribute to immersive storytelling experiences (Verma, 2019; Mildorf, 2019).

Further investigations explore specific genres, such as Gothic fiction, to uncover how ambient soundscapes shape narrative atmospheres and reader perceptions (Guhr and Algee-Hewitt, 2024). Fine-tuned BERT models successfully detect and analyze word-level sound indicators across literary texts. This interdisciplinary approach not only enhances our understanding of sound's role in genre classification but also sheds light on broader narrative techniques and reader engagement strategies.

3 Scandinavian Literary Soundscapes

Studies in sensory history, and particularly *soundscape* studies, stress the importance of the sonic environments in which people live and operate. The importance of these soundscapes is based on the premise that the sounds one hears in a given place are as distinctive and as important as the things one sees there (Birdsall, 2012). 19th century urbanized settings presaged a 'sound revolution,' where novel individual sounds and soundscapes rapidly emerged due to new industries, machinery, means of transportation, road surfaces, and the like (Parby, 2021). Simultaneously a fundamental shift in people's relationship with sound and noise took place, leading some scholars to characterize the 19th and early 20th centuries as a particular auscultative age—an era devoted to sonic experiences, to 'close listening' and to auscultation "not only in the medical sense initiated by the stethoscope [. . .] and perfected by the microphone [. . .] but also in the sense of careful listening to a world at large." (Picker, 2003).

In a Scandinavian context, these sensory events are related to the "modern breakthrough." Copenhagen changed radically during this time, with an an explosion in its population, and the introduction of new technologies and infrastructures. As Parby (2021) has shown, these developments led to the emergence of new soundscapes, which were then incorporated into contemporary fiction. Simultaneously, authors developed new realist literary techniques to give a fuller account of the material world and to represent it with verisimilitude and detail (Bjerring-Hansen and Wilkens, 2023).

4 Defining Noise

In Sound Studies, the definition and phenomenological demarcation of noise has been a point of considerable discussion. Etymologically, the word 'noise' is rooted in Latin 'nausea' that encompasses seasickness and nausea, and refers to sounds that are perceived as excessive, incoherent, confused, or twisted. One of the founders of Sound Studies, Raymond Murray Schafer, proposed a clear dichotomy between natural and man-made noises from, for instance, industrial activity and traffic (Schafer, 1993). This position has later been criticized for being a far too normative and rigid division, where man-made noise almost by definition is perceived as negative and natural sound events such as thunder claps as positive (Kelman, 2010).

Some more subject-sensitive, less normative and, not least, operational definitions, which we rely on, have been suggested by David Novak who defines noise as "deviant sonic behaviour" (Novak and Sakakeeny, 2015), and Peter Bailey who defines it as 'sound out of place' paraphrasing anthropologist Mary Douglas' classic definition of dirt as 'matter out of place' (Bailey, 1996).

In order to develop an operational conceptual framework, we use a very basic definition of noise as "silence-breaking"; this characterization means that noise can include moderate sounds such as whispering or mumbling voices as long as they are noted as sonic events in the text. It also means that we include neutral sonic phenomena, such as this description of factory whistles:

> The steam pipes sounded from all the factories, it was 8 o'clock. And I had to leave. (transl.)

The quote says nothing about whether the whistles are perceived as noise by the character or narrator. But, obviously, the sound distinctively breaks a silence.

Based on our knowledge of industrialization and urbanization, as discussed in international Sound Studies research and reflected in literary sources, our hypothesis in the following is that during the latter part 19th century the general noise levels increased, correlated with an increasing sensitivity towards noise, in a development more and more dominated by non-human noise sources.

5 An Annotated Dataset of Noise in 19th Century Scandinavian Literature

We introduce a framework for noise detection and categorization, and apply it to a corpus of Scandinavian literature, creating a new annotated dataset.

5.1 Main Corpus

For our main target data, we rely on the MeMo corpus (Bjerring-Hansen et al., 2022), comprising 859 Danish and Norwegian novels spanning the last 30 years of the 19th century, with more than 64 million tokens. We refer to this corpus as the 'main corpus'. The corpus is a rich and diverse collection of texts that provides valuable insights into the representations of noise and sound during the period under investigation. Table 1 shows statistical information about the corpus. We segment the corpus into paragraphs and split them into 50-word segments if they exceed 50 words.

Total novels	859
Total segments	1,936,527
Total words	64,227,927
Average segments per novel	2,254
Average words per novel	74,771
Average words per segment	33

Table 1: MeMo corpus statistics.

5.2 Noise Detection Dataset

To construct a dataset of text segments annotated for whether they contain noise or not, we combine a selection of hand-picked segments by a historian expert, with a topic-based search approach to enrich the dataset.

Segments Extraction. We apply BERTopic (Grootendorst, 2022), a powerful topic modeling technique that enables us to cluster millions of text segments from the MeMo corpus into a concise set of topics. By doing so, we aim to distill vast amounts of textual data into manageable thematic clusters, facilitating subsequent analysis. Following the topic modeling phase, we filter the generated topics, focusing our attention on those most relevant to the concept of noise. These selected topics serve as a foundation for further exploration, guiding us in identifying and annotating text segments specifically related to noise. Table 2 shows the list of topics that are used to extract the text segments from the main corpus. We remain with 5,700 text segments related to noise topics, which are then carefully annotated by experts.

Annotation Guidelines. The annotation involves two of the authors, native Danish speakers, a historian with special interests in urban and social history as well as a literary scholar familiar with the social conditions of 19th century literature, who classified segments into two categories: noise and non-noise. This differentiation is by no means trivial.

With the conceptual distinctions and demarcations in §4 in mind, the annotation of noise and non-

Topic	Translation
lyd_lyden_støj_larm	sound_the sound_noise_noise
lydbevægelse_vindtrækket_sporvognens_ringlende	sound movement_the draft_the streetcar's_ringing
støy_hviskedee_blef_hviskedees	noise_whispered_was_whispered
lyd_betjening_lydløs_værkstedet	sound_operation_silent_the workshop
lyd_lyde_lydt_fest	sound_sounds_sounded_party
overraskelsesskrig_højtalende_lyd_hørtes	surprise scream_loudly_sound_was heard
klinkede_klink_klinke_klinkes	clinked_clink_clink_clinked
gadetummelen_færdselens_støj_vognrullen	street noise_the traffic's_noise_the wagon's rolling
skriger_skrige_skrigene_øses	screams_scream_the screams_is scooped
højt_højlydt_snakker_snak	loud_loudly_talks_talk
råbteredando_råber_skreget_opmærksomme	shouting_shouts_the scream_attentive
skriger_råber_skreg_hvorfor	screams_shouts_screamed_why
råber_råbes_råpte_mode	shouts_is shouted_shouted_fashion
skreg_skrek_skrige_skriker	screamed_screamed_scream_screams
signalet_signal_signaler_udskælden	the signal_signal_signals_scolding
råbte_råbt_borde_skreg	shouted_shouted_tables_screamed
fløjten_fløjte_fløjter_fløjtede	the whistle_whistle_whistles_whistled

Table 2: List of selected topics with English translations

noise text segments is carried out on the basis of the following, minimalist and pragmatic guidelines, respecting the principle that clear and simple instructions are crucial for obtaining high-quality annotations (Mohammad, 2016), while also acknowledging the intricacies that the analysis of literary texts based on small fragmentary segments raises.

1. Based on our definition of noise, the text segments are labeled either '1' (positive) or '0' (negative). Our focal point is that we would rather narrow the focus later on (through additional annotation of positive cases, metadata-filtering, or NLP measures) than exclude specific types of noise. Along the way, our definition becomes even richer, as we realize that music should also be included, as music is often interwoven with other types of sonic events, as we see in this example:

   > The orchestra began playing a French folk tune ... and the stormy applause and applause clapping eventually faded away. (transl.)

2. Only the segment in question should be considered. Contextualisation and 'guessing' on what might go on before or after the segment were ruled out. In cases of doubt, the label should be '0'. Example:

   > But after the noisy scenes he had caused, came the lethargy that always follows the performance of a great tragic role. (transl.)

3. Negated noise should be filtered out, e.g,

   > The gardens alongside the houses looked very well on this summer evening. There was no noise or disturbance, only a couple of children playing across the street, but they did not chatter; even their play was in keeping with the tone of the evening. (transl.)

4. The same goes for other pseudo-relevant segments detected by the topic modeling algorithm, including noise metaphors and similes (prompted by words like 'as' or 'like'), as we see in this example:

   > She felt engulfed by a buzzing electric current. (transl.)

Annotation Results. Our hand-picked selection of noise segments include 217 positive examples, manually curated from various 19th century sources (memoirs, essays and fictional works). As for the segments obtained with the topic-based search, out of the pool of 5700 segments, 337 are deemed noise-related while the remaining segments are annotated as non-noise.

In our endeavor to encompass the entirety of the target corpus, we turn our attention to the remaining 5,365 segments, considering them as "non-noise" segments. To ensure the dataset's comprehensiveness and diversity, we supplement these non-noise segments with an additional 5,000 segments randomly selected from various non-noise topics.

We randomly sample 175 segments from the noise-related topics to serve as our testing set, each annotated independently by our two annotators, to evaluate annotation consistency and assess model performance. The resulting Cohen's Kappa value, calculated to measure inter-annotator agreement, yielded a score of 0.85, indicating a high level of agreement between the annotations.

5.3 Noise Categorization Dataset

Fine-grained categorization of noise-related segments is essential in the context of classifying textual noise extracted from literary texts. This approach enables a nuanced understanding of the diverse forms of noise present within the textual corpus, including but not limited to, linguistic anomalies, contextual inconsistencies, and stylistic irregularities. By classifying noise into categories such as human noise, mechanical noise, and other types of textual disturbances, one can distinguish specific sources of interference more accurately, reflecting the intricacies inherent in literary compositions.

Annotation Guidelines. To get closer to an understanding of sound as a cultural phenomenon as reflected in literary works, we perform another round of annotation. Although we have several specific research interests related to sonic developments in the 19th century, in order to (a) reduce the number of axes in the annotation, which might have negative consequences for the predictive power of the model, and (b) produce a more broadly useful dataset, we choose to prioritize one aspect of the noise segments, namely the *sound source*. In doing so, we have disregarded features, which, in a future, extended pipeline, may come into play, not least *time* (traditional or modern sound?) and *place* (rural or urban sound?).

For this round of annotation, we merge the noise-related segments from the previous dataset and additional segments from the MeMo corpus after the prediction of noise and non-noise classes for each segment in the corpus as shown in Figure 1. Then, the (same two) annotators classified text segments into the following categories: Non-human made noise (T), Human-made noise (H), Undefined noise (N), and Music (M), following these criteria to ensure an accurate and consistent categorization:

1. Non-human made noise encompasses any noise not produced by humans, ranging from machine-produced sonic events (such as steam engines, trams, telephones etc.) to natural ones (caused by wind, rain, animals etc.).

2. Human-made noise includes any noise resulting from human activities (such as footsteps, conversation, yelling, booing etc.).

3. Undefined noise is the appropriate label when the noise source is unknown or unclear, as in this example where the noise is abstract and generic:

   > Outside, the music stopped and Madsen's voice was barely audible through the noise. (transl.)

4. Music is a special category in relation to both the general ontology: noise yes/no? (see above) and to the specific categorization. Since it is futile to determine whether a rattling sound is produced by the violin player or his instrument, we decided to give music a label of its own.

5. Often there is a mix of sound sources in the individual text segments, as here (non-human made, human made, music):

   > From time to time there were snatches of a loud violin's dance tunes. Lonely cabs rumbled through the street, with snow-damped wheels and a few swishing whistles that made a couple of heads turn in the window. Every now and then a streetcar threw its jingle of bells and chimes into the whispers and murmurs of conversation. (transl.)

   Here, non-human made noise dominates, so the label is 'T'. In other cases, the categorization of mixed segments is based on a more uncertain basis and is open to interpretation.

Annotation Results. Following the annotation process, we have a total of 1,874 text segments annotated by two independent annotators. Annotated data statistics are presented in Table 3. The training set, encompassing ∼91% of the total annotations, consists of 1,699 segments, while the test set, comprising ∼9%, consists of 175 segments. After the removal of non-noise segments, the total number of segments in the dataset is 1,244. Notably, both annotators annotate all segments within the test set, ensuring comprehensive coverage and reliability. Our obtained Cohen's Kappa value of ∼0.81 demonstrates a substantial agreement level, surpassing chance expectations. They exchange opinions on the interpretation of borderline cases, especially regarding segments including a multitude of noise sources, in order to establish a common understanding of the different noise categories. As a result of these initial considerations we also decided to give music its own category. This result underscores the robust and accurate classification of the data, reflecting strong and reliable consistency in the annotations provided by both annotators. Note that in the training set, each annotator individually annotates half of the segments, maintaining an equitable distribution to uphold annotation quality and consistency across the dataset.

Non-human	Human-made	Undefined	Music
513 (40%)	424 (33%)	56 (4%)	269 (21%)

Table 3: Noise categorization annotated data statistics.

6 Experiments and Results

In this section, we describe the selection of pre-trained language models as well as the classification experiments on the noise detection and noise categorization datasets.

6.1 Pre-trained Language Models

In this subsection, we outline the models evaluated in our noise detection and categorization classification experiments using supervised fine-tuning methods. Importantly, all models are selected based on their performance evaluated on Danish and Norwegian literary benchmark datasets (Al-Laith et al., 2024) and ScandEval[1] (Nielsen, 2023), even though these models had not been trained primarily on historical Danish or Norwegian.

DanskBERT. DanskBERT[2], a top-performing Danish language model noted for its success on the ScandEval benchmark (Snæbjarnarson et al., 2023), is based on the XLM-RoBERTa architecture and trained on the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021). It features 24 layers, a hidden dimension of 1024, 16 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 (Enevoldsen et al., 2023) 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.[3]

MeMo-BERT-03. Developed by continuing the pre-training of the Transformer PLM DanskBERT (Al-Laith et al., 2024).[4] 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 (Al-Laith et al., 2024).

NB-BERT-base. A general-purpose BERT-base model was developed using the extensive digital collection at the National Library of Norway (Kummervold et al., 2021).[5] 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.

6.2 Experimental Setup

In this section, we outline the experimental setup employed for the supervised classification tasks focused on both noise detection and noise categorization. Our experiments involve fine-tuning several pre-trained language models on the fine-grained datasets. All layers of the selected models were actively trained to optimize performance. The details of the dataset and the models used are described below. For the training procedure, the experiments involve fine-tuning BERT models on the dataset using a batch size of 32, training for 20 epochs with the AdamW optimizer at a learning rate of 10^{-3}. During training, we monitored both training and validation losses to assess model convergence and prevent overfitting. For evaluation, we employed the F1-score metric due to its ability to balance precision and recall, particularly effective for tasks with imbalanced datasets like noise detection and categorization. The performance of each model was evaluated on both validation and test sets, ensuring the robustness and generalizability of the models across different datasets and epochs.

[1] https://scandeval.com/

[2] https://huggingface.co/vesteinn/DanskBERT

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

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

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

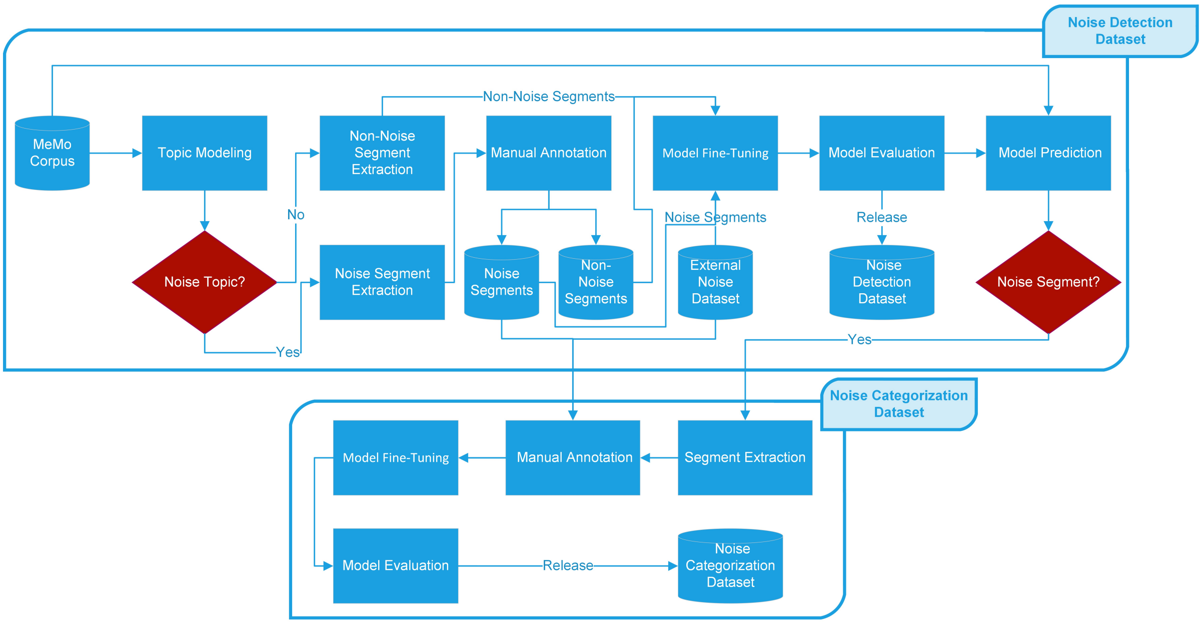

Figure 1: Noise Datasets Creation Flowchart.

6.3 Noise Detection Experiments

It is important to note the deliberate imbalance within the dataset, where only ∼5% of the annotated segments are noise-related. By favoring a higher representation of non-noise segments, we aim to bias our model toward accurately identifying and capturing instances of noise within the data. This approach is designed to enhance the model's sensitivity to noise while maintaining robustness in its classification capabilities.

Fine-tuning the PLMs on the noise detection task results in notable performance variations (Table 4). DanskBERT achieves a validation accuracy of 0.89 and a test accuracy of 0.83, indicating robust performance across unseen data. MeMo-BERT-03 demonstrated the highest validation accuracy at 0.90, although its test accuracy slightly decreased to 0.80. In contrast, DFM (Large) exhibited a validation accuracy of 0.81, dropping significantly to 0.55 on the test set, suggesting potential overfitting or limited generalizability. NB-BERT-base achieved consistent results with a validation accuracy of 0.88 and a test accuracy of 0.76, indicating reliable performance across both validation and test datasets. These results highlight the effectiveness of fine-tuned BERT variants, especially MeMo-BERT-03 and DanskBERT, in accurately detecting noise within textual data, while emphasizing the importance of robust evaluation across multiple models and datasets. Table 4 also shows a detailed of result of precision and recall of positive class.

6.4 Noise Categorization Experiments

The dataset comprises 1,244 text segments, divided into training, validation, and testing sets for model development and evaluation. The training set includes 961 examples, constituting ∼77% of the dataset, while the validation set, used for hyperparameter selection, consists of 178 samples, representing ∼14% of the total. The testing set, for the final model evaluation, contains 105 examples, or ∼9% of the dataset. Annotations for the training and validation sets were made by a single expert. For the testing set, only segments where both experts agreed on the annotations were retained, discarding those with conflicting annotations. We use the weighted average F1-score as the evaluation metric. Notably, MeMo-BERT-03 achieved the highest F1-score of 83% on the validation set, while the DanskBERT model achieved the highest F1-score of 83% on the test set. Table 4 shows detailed results for each model.

7 Diachronic Analysis of Noise Segments

Having trained accurate noise detection and categorization classifiers, we use the best ones (DanskBERT fine-tuned on the two tasks respectively) to predict labels for all segments in the entire MeMo corpus. We then quantify the frequency of the occurrence of noise over time in the corpus, as well as the distribution of the different categories.

	Detection				Categorization	
	Valid.	Test			Valid.	Test
Model	F1	F1	Precision	Recall	F1	F1
DanskBERT	0.89	**0.83**	0.82	0.59	0.81	**0.83**
DFM (Large)	0.81	0.55	0.36	0.85	0.82	0.79
MeMo-BERT-03	**0.90**	0.80	0.60	0.80	**0.83**	0.80
NB-BERT-base	0.88	0.76	0.59	0.52	0.81	0.81

Table 4: Validation and test F1-Score results of fine-tuning the selected models on the noise detection and categorization datasets. The Precision and Recall for the positive (Noise) class of noise detection testing set.

7.1 Noise Occurrences over Time

After fine-tuning multiple pre-trained PLMs, DanskBERT emerged as the top performer among the four models evaluated, and we selected it for predicting noise and non-noise classes across all segments in the main corpus. Notably, out of 1.9 million segments in the corpus, 220,378 were predicted with the noise class. Figure 2 shows the proportion of noise segments over the years.

A trend of rising noise levels in the novels is clear: From the 1870s to the 1890s there is a more than 50 % relative increase of noise (followed by a slight decline or a plateau by the end of the decade).

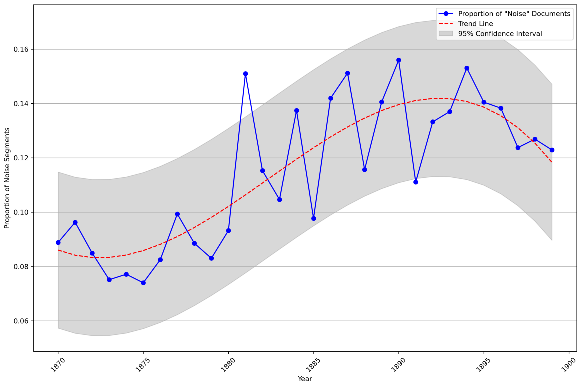

Figure 2: Proportion of Noise Segments Over Years.

7.2 Noise Categories over Time

Applying the best classifier for the second task (DanskBERT fine-tuned on noise categorization), we predict noise categories for all positive predictions in the corpus from the previous step. Figure 3 shows the frequency of the noise categories.

The analysis indicates a stable distribution of the different kinds of noise without significant fluctuations.

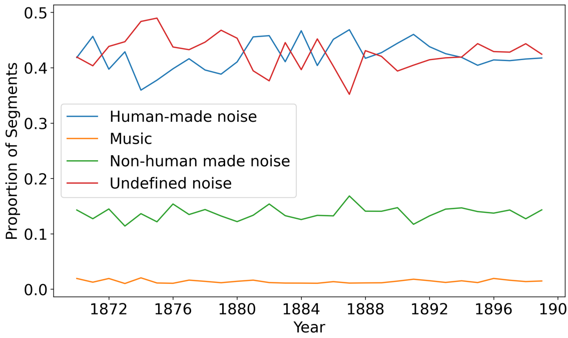

Figure 3: Frequency of Noise Categories over Time.

8 Discussion

Annotation nuances. Given the complex and slippery character of noise as a cultural phenomenon, we chose to disregard both spatial aspects (where?) and historical aspects (when?) related to it. For pragmatic reasons, we decided on a basic categorization by annotating noise classes (based on noise sources). There were challenges in making categorical decisions about specific noise events in the texts, for instance when there was a mix of noise sources simultaneously at play and, not least, in relation to our crucial distinction between non-human made and human made noise, according to which in the former category, it is humans themselves who generate sound (with their voice or body), while in the latter, technology is perceived as an agent and humans as mere operators (e.g. of a ringing church bell or a screeching streetcar). Our categories proved to be operational, but they are by no means watertight. We have learned more about noise as a historical concept, but also about its dynamic way of manifesting itself in different contexts and literary representations.

Model performance. Interestingly, while MeMo-BERT-03 achieved the highest validation scores, DanskBERT outperformed it on the test set for both noise detection and categorization (Table 4), suggesting it is more capable at generalizing to unseen segments. While the former is designed to be better attuned to nuances in the historical corpus, in some cases, the latter (which was trained on a diverse corpus consisting mostly of modern Danish) might be better at detecting modernity signals, which are the focus of our annotation framework.

Noise trends. Our results confirm our hypothesis that in the last three decades of the 19th century a rise of general noise levels as well as an increase in preoccupation with noise are reflected in the novels of the time. The upward trend is clear. It does, however, flatten or fall by the end of the period. This is hardly due to less noise, but rather to the fact that the authors and their characters are less preoccupied with it. Our hypothesis, which must be supported by close inspection and reading, is that noise is taken for granted or implied on the brink of the 20th century. In contrast to our initial hypothesis, human-made noise remains at a relatively high level throughout the period. This fits well with observations from the larger European metropolises like Paris, London and Madrid, where the policing of human noise sources remain significant, whereas industrial sounds tends to be evaluated positively and is not the focus of anti-noise campaigns until the early 20th century. The results from our noise categorization do however call for further time- and place-attentive investigation.

9 Conclusion

We presented a framework for detecting and categorizing noise in literary texts and demonstrated its usefulness in the MeMo corpus. Using topic modeling and fine-tuned BERT-based models, we extracted and analyzed relevant text segments, providing new insights into the cultural and social transformations reflected in the soundscapes of the Scandinavian “Modern Breakthrough” period. Our study demonstrates that literary perceptions of noise can be effectively tracked and categorized, revealing significant patterns and trends. We have been able to add new perspectives on the interplay between literature and cultural history – and to empirically underpin hypotheses of the 19th century as a particular auscultative era.

Future work will extend this framework to explore the impact of industrialization, examining how technological advancements and urbanization influenced literary soundscapes. We will also investigate the spatial dimensions of noise, contrasting rural and urban settings. Further, we would like to do comparative analysis on other datasets to situate our study in a broader context, such as contemporary civil complaints (as documented in the City Archives), and/or a corpus of modern Danish novels (from The World Literature Data Collective). Additionally, we plan to analyze the lexical diversity in the terms used to portray noise, to get a better understanding of the psychological and cognitive aspects of the increased awareness of noise in the early phases of urbanization and industrialization.

Limitations

Despite the strengths of our framework, there are several limitations to consider. First, the focus on Danish and Norwegian literature from a specific historical period may limit the generalizability of our findings to other linguistic and cultural contexts. Second, the accuracy of our noise detection and categorization relies heavily on the quality of the annotations and the pre-trained language models, which may not capture all nuances of noise representation in literary texts. Third, our current analysis does not account for the broader contextual elements surrounding noise occurrences, such as narrative structure or character perspectives, which could provide a deeper understanding of the literary soundscapes. Finally, while our framework demonstrates promising results, further validation across diverse datasets and more complex noise categorization schemes is necessary to fully establish its robustness and applicability.

Ethics Statement

The MeMo corpus, which we use, is released under the Creative Commons Attribution 4.0 International license.

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.

Peter Bailey. 1996. Breaking the sound barrier: A historian listens to noise. *Body & Society*, 2(2):49–66.

Carolyn Birdsall. 2012. *Nazi soundscapes: sound, technology and urban space in Germany, 1933-1945*. Amsterdam University Press.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts. a heuristic procedure for correcting OCR data.

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.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Jason Dou, Ni Sun, and Xiaojun Zou. 2016. " draw my topics": Find desired topics fast from large scale of corpus. *arXiv preprint arXiv:1602.01428*.

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.

Angela Frattarola. 2018. *Modernist soundscapes: Auditory technology and the novel*. University Press of Florida.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Svenja Guhr and Mark Algee-Hewitt. 2024. What's that scary sound? ambient sound in gothic fiction. *Journal of Computational Literary Studies*, 2(1).

Thomas N Haider. 2019. Diachronic topics in new high german poetry. *arXiv preprint arXiv:1909.11189*.

Gerhard Heyer, Cathleen Kantner, Andreas Niekler, Max Overbeck, and Gregor Wiedemann. 2017. Modeling the dynamics of domain specific terminology in diachronic corpora. *arXiv preprint arXiv:1707.03255*.

Ari Y Kelman. 2010. Rethinking the soundscape: A critical genealogy of a key term in sound studies. *The senses and society*, 5(2):212–234.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Nora Elisabeth Lambrecht. 2017. *The Art of Noise: Literature and Disturbance 1900-1940*. Ph.D. thesis, Johns Hopkins University.

Jani Marjanen, Elaine Zosa, Simon Hengchen, Lidia Pivovarova, and Mikko Tolonen. 2020. Topic modelling discourse dynamics in historical newspapers. *arXiv preprint arXiv:2011.10428*.

Jarmila Mildorf. 2019. Can sounds narrate? prosody in sound poetry performance. *CounterText*, 5(3):294–311.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Dan Saattrup Nielsen. 2023. Scandeval: A benchmark for scandinavian natural language processing. *arXiv preprint arXiv:2304.00906*.

David Novak and Matt Sakakeeny. 2015. *Keywords in sound*. Duke University Press.

Jakob Ingemann Parby. 2021. Fremskridtets lyd? lydrevolutionen og håndteringen af støj under københavns industrialisering ca. 1850-1910. *Kulturstudier*, 12(2):41–71.

John M. Picker. 2003. 3INTRODUCTION: The Tramp of a Fly's Footstep. In *Victorian Soundscapes*. Oxford University Press.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

R Murray Schafer. 1993. *The soundscape: Our sonic environment and the tuning of the world*. Simon and Schuster.

Hale Sirin and Tom Lippincott. 2024. Dynamic embedded topic models and change-point detection for exploring literary-historical hypotheses. *arXiv preprint arXiv:2401.13905*.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. 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.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Timothy R Tangherlini and Peter Leonard. 2013. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, 41(6):725–749.

Leah Hutchison Toth. 2016. Resonant texts: Sound, noise, and technology in modern literature.

Neil Verma. 2019. *Theater of the mind: imagination, aesthetics, and American radio drama*. University of Chicago Press.