Evaluating Embedding Models on Danish Historical Newspapers: A Corpus and Benchmark Resource

Alie Lassche*, Pascale Feldkamp*, Yuri Bizzoni*, Katrine Baunvig†, Kristoffer Nielbo*, Johan Heinsen‡

*Center for Humanities Computing, Aarhus University, †Center for Grundtvig Studies, Aarhus University, ‡MASSHINE, Aalborg University

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

Abstract

We present an enriched dataset of almost five million Danish historical newspaper articles from the late seventeenth to nineteenth century, augmented with semantic embeddings and an annotated subset, to enable semi-automated classification as well as thematic and linguistic exploration. Through three historical benchmark tasks that evaluate the performance of Danish and multilingual embedding models on this historical Danish corpus, we discuss how the choice for an embedding model depends on the type of task, and enrich our corpus with embeddings from the overall best performing model. As a showcase experiment, we look at the distribution of article categories in the three subgenres that can be observed in the corpus. This experiment highlights the corpus and article-level embeddings' potential for further exploration and analysis of the Danish historical mediascape. The resource is freely available for research use and aims to foster reproducible, data-driven studies of language and culture in the Danish nineteenth century.

Keywords: historical corpus, Danish, historical newspapers, OCR, embeddings, diachronic linguistics

1. Introduction

In the rapidly advancing field of making textual corpora digitally accessible, it is often modern corpora of high-resource languages taking the lead. Such corpora – consisting of neatly typeset pages in legible fonts – are relatively straightforward to make available in full text. Their enrichment with numerical representations such as sentence or document embeddings, and their annotation, is facilitated by the multitude of transformer models trained on these languages and fine-tuned for these tasks, and covering these time periods.

This stands in contrast to historical corpora of low-resource languages, which tend to lag behind. This is partly due to the challenges of making such texts available in clean, full-text form: early modern prints often feature worn type, irregular page layouts, and typographic conventions that resist automated processing. Difficult fonts, substantial spelling variation, and the absence of standardized stylistic or grammatical norms further complicate OCR accuracy (Burchardt, 2024).

In addition, while embedding benchmarks provide valuable insights into which models perform best on a given task in contemporary settings, these evaluations are almost exclusively carried out on modern corpora and domains (Muennighoff et al., 2023). As a result, we know little about how well these models transfer to historical texts, where the linguistic signal is shaped by orthographic instability, genre conventions, and cultural-historical specificities that are absent from present-day benchmarks. This lack of tailored evaluation makes it uncertain whether strong performance on modern benchmarks translates into reliable representations for historical corpora and tasks.

Together, these factors make downstream analysis of historical texts more challenging. Robust resources for such corpora therefore require not only more labor-intensive preprocessing but also methodological solutions that are sensitive to the linguistic and material conditions of historical print culture (Mordell, 2019).

Against this backdrop, we present a large-scale corpus of eighteenth- and nineteenth-century Danish newspapers and periodicals: a resource that fills precisely this gap by offering clean (enough) OCR, rich metadata, and new possibilities for computational exploration of the Danish language and pre-modern culture. In addition, we introduce three historical benchmark tasks that evaluate the performance of Danish and multilingual embedding models on this historical Danish corpus, thereby offering a first step toward assessing how well contemporary embedding approaches transfer to historical language data.[1]

This paper has the following structure: Section 2 reviews related work on historical Danish newspa-

[1]To ensure reproducibility, our code is available via https://github.com/centre-for-humanities-computing/danish-newspaper-embs. The enriched corpus, including embeddings, is available via https://huggingface.co/datasets/chcaa/eno-embs-old-news.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3577–3589
11–16 May 2026.

pers, the digitization of newspaper corpora in other languages, and the development of the Scandinavian embedding benchmark. Section 3 introduces the characteristics of the corpus presented in this paper. Our methodological pipeline is described in Section 4; it comprises OCR processing, followed by three benchmark tasks. Section 5 discusses the results of the benchmark tasks (an article category classification task, a fiction/non-fiction classification task, and a feuilleton clustering task), along with a discussion on how the final model in this task was chosen. We finish this section with an experiment in which we showcase the potential of the article-level embeddings, by looking at the average category distributions per newspaper group in the corpus. Section 6 concludes the paper, with suggesting some directions in which future work should go.

2. Related Work

In Danish and Norwegian contexts, historical newspapers from before the twentieth century have been largely neglected by researchers. The few studies that do exist can be divided into two categories. On the one hand, earlier work has focused on paratextual metadata such as print runs and sales numbers (Kjærgaard, 1989; Horstbøll, 1999). These studies have provided valuable insights into, for example, the geographical distribution of newspapers. On the other hand, more recent research has taken a quantitative approach to specific case studies ranging from communal singing in the nineteenth and twentieth century to wanted notices in a nineteenth century newspaper from Copenhagen, and from the demi-goddess Dana in a Danish nationalist context to representations of Lourdes in Denmark, using newspapers as their main source material (Agersnap et al., 2025; Heinsen and Birkemose, 2023; Baunvig, 2021, 2023). Yet despite these contributions, we still lack *large-scale* studies that systematically analyze the *content* of the newspapers. As a language resource, too, the material remains underused, primarily due to limited accessibility of the texts. Although Danish and Norwegian libraries hold well-preserved digital newspaper collections, their quality makes it difficult for researchers to work with these corpora (see also section 3).

By contrast, in other countries the large-scale digitization of historical newspapers has long been a priority. In the past years, collections of British, Finnish, Swedish, German, French, Luxembourgish, and Spanish newspapers – to name a few – have become available. These initiatives go beyond simply putting digitized newspapers online: they also aim to improve OCR quality (Koistinen et al., 2017; Brandt Skelbye and Dannélls, 2021; Thomas et al., 2024; Löfgren and Dannélls, 2024) and to develop advanced methods for information retrieval (Ehrmann et al., 2020), named entity recognition (Ruokolainen and Kettunen; Tudor et al., 2025), and even irony detection (Cohen et al., 2025). The goal is to enable any interested user to work with the material more effectively.

The dataset presented in this study represents a first step toward making Danish historical newspapers accessible for large-scale computational research. To this end, we enrich the newspaper articles with text embeddings. For modern Danish corpora, the Scandinavian Embedding Benchmark (SEB) has been a pivotal initiative in this context (Enevoldsen et al., 2024). SEB provides a comprehensive framework for evaluating text embeddings for Scandinavian languages across five task types: classification, clustering, retrieval, bitext mining, and linguistic acceptability. It currently includes more than 50 models – both mono- and multilingual – allowing users to identify which models perform best for their specific tasks on Scandinavian text corpora. However, most SEB tasks rely on modern language data, making it difficult or even impossible to directly extrapolate the results to historical corpora. Therefore, in this paper, we introduce three custom benchmark tasks designed to identify the most suitable model for our corpus of historical Danish newspapers.

3. Corpus

Originating from the early seventeenth century, the newspaper became an everyday medium bringing international and national news alongside advertisements and announcements to a broad reading public from the second half of the eighteenth century (Weber, 2006; Pettegree, 2014). In addition to simply transmitting information to its readers, early newspapers also created a shared forum of discussion in the public sphere, thus enabling critical debate (Habermas, 1989). Moreover, the daily act of reading the same printed matter helped people imagine themselves as part of a common nation (Anderson, 2016).

In the conglomerate state of Denmark-Norway – united until 1814 – newspapers went from being a medium for Copenhagen's political and merchant elites to being published in substantial numbers across all major provinces (Kjærgaard, 1989). This is reflected in an expansion of the contents. The same edition that told of major world events, natural disasters, or royal pageantry might eventually also contain the advertisement of a rural servant looking for a household to serve in town or that of a farmer looking for his missing horse (Søllinge and Thomsen, 1988). By the early nineteenth century, more and more papers carried serialized fictional contents too (Lehrmann, 2018; Feldkamp et al., 2025).

This breadth of perspectives makes the newspapers a unique source for historical research into social or cultural dynamics over time. Yet, newspapers also offer a diversity of linguistic representations of a range of human and natural phenomena. Thus, to anyone working on language use in diachronic perspective they are highly valuable.

The corpus we present here consists of 28 Danish historical newspapers and periodicals, published between 1666 and 1850 in Denmark-Norway. The newspapers are kept in extraordinarily well-preserved collections housed by the national libraries of Denmark and Norway.[2] The editions are included as they exist in the library repositories. This means that any gaps in the series reflect lost materials.[3] While most series are complete or near-complete, the inclusion of various appendices is highly uneven. Such appendices were typically smaller periodicals from the same publisher, but published at slower rates than the main paper. For example, towards the end of the period, publishers often published minutes from the estates assemblies (*stænderforsamlinger*) and local citizen councils (*borgerrepræsentationer*) and bundled them with the regular newspapers (Jørgensen, 1956; Friisberg, 2013; Jensen et al., 1934). Various subscription-based fundraising schemes are another common element among the appendices, as well as more standard extra editions of the regular newspapers. In the rare cases where appendices appear undated in the collections, they are left out.

Based on domain knowledge, we distinguish three subgenres within the corpus, which mirror shifts in Danish press culture: *(1)* early newspapers (published before 1800) followed a European style and focused largely on royal and diplomatic news; *(2)* local and national papers – emerging in the later eighteenth century – mixed local notices, advertisements, and practical information, reaching a much broader reading public; and *(3)* opinion-based papers, predominantly from the nineteenth century, moved toward commentary and political agitation around the coming constitution. A compact overview of the newspapers in the corpus, with descriptive statistics on these subgenres, is included in Table 1. In total, we are dealing with a corpus of almost five million articles, consisting of 473 million words. Their distribution over time is visualized in Figure 1.

While Copenhagen is up until today considered the urban center of the Kingdom of Denmark, the

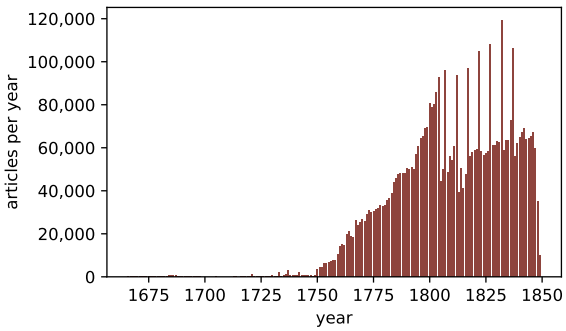

Figure 1: Distribution of the corpus, expressed in number of articles per year.

population of the conglomerate state Denmark-Norway lived first and foremost in the periphery during the period that is covered in our corpus. This geographical spread is reflected in the corpus presented here: although it includes five newspapers that were published and distributed in Copenhagen, newspapers from the peripheral towns and villages of Aalborg, Aarhus, Ribe, Odense, Viborg, Maribo, Slagelse and Thisted are also represented, as well as the titles from the nowadays Norwegian cities Trondheim and Bergen.

Unfortunately, the newspaper collections have lacked quality in their transformation from paper to digital text. Traditional OCR technologies have struggled with both language, layouts, and letterforms. The material is set in *fraktur* types, cheaply printed on thin, now deteriorated, paper sheets with relatively complicated layouts. The approximate word-level accuracy of the OCR was below 50 percent on the Danish collection.[4] It appears marginally better for the Norwegian one, probably because it was photographed more recently.

4. Method

We use a three-step pipeline to create our digitized corpus and enrich it with embeddings:

1. OCR processing the corpus;
2. Benchmark task I: article category classification;
3. Benchmark task II: fiction/non-fiction classification;
4. Benchmark task III: feuilleton clustering.

[2] https://www.mediestream.dk and https://www.nb.no/search?mediatype=aviser.

[3] The one exception to this is *Københavns Adresseavis*, which from 1804 onwards is sampled at five year intervals, due to its size and uniformity of contents.

[4] This estimate is based on averaging the word accuracy scores provided when the series are accessed via the Mediestream API.

Group	Period	Editions	Articles	Words
Early newspapers	1666–1798	3,261	71,186	6,972,443
Local/National papers*	1749–1848	77,778	3,941,251	436,083,783
Opinion-based papers	1798–1844	12,662	124,176	29,576,426
Outlier: *Adresseavis for Børn***	1779–1782	206	1,471	504,229
Total		93,908	4,898,084	473,136,881

Table 1: Compact overview of newspapers in the corpus, grouped by historical and thematic type. *While newspapers in this category all feature a significant section devoted to national news, their reporting is primarily local in scope. The exception is the Copenhagen-based *Berlingske Tidende*, whose coverage is truly national. **The *Adresseavis for Børn* is listed as an outlier due to its unique target audience – children.

4.1. OCR processing the corpus

Due to the poor OCR quality, the newspapers in this corpus had to undergo a re-digitization process. This was carried out by historians at Aalborg University via the Transkribus platform, a tool originally designed for Handwritten Text Recognition (HTR) but currently the most accessible and popular platform for both HTR and OCR tasks (Kahle et al., 2017). It involved training custom layout and text recognition models tailored to the collections in question (Heinsen and Bøgeskov, 2025).[5]

A range of layout models were designed to account for diversity and evolutions in newspaper layouts. Consistently, the main challenge was to identify separate columns in layouts with only hairline vertical separators between text blocks. The fraktur type itself had less variation across the corpus, and could therefore be handled without training separate models for different types. Instead, the text recognition model was improved iteratively as new typographical variants appeared in the corpus. The model was built from an initial base of 250,000 words stemming from the period 1770–1825. Training data for the final version eventually grew to include about 420,000 manually transcribed words, spanning multiple titles from across the entire period. The model was built using Transkribus' implementation of the PyLaia framework.

Since photographing the pages over again would have been prohibitively costly, both layout and text recognition models were designed to process the images as they already existed in their respective library collections. For the vast majority, stemming from Denmark's Royal Library, this meant working from scans of microfilm-copies of the original newspapers. In some cases, these microfilm-copies date as far back as the 1950s, and the resulting grayscale images often suffer from poor contrast – a characteristic that had proven a major obstacle for previous attempts at digitization. However, by tailoring the text recognition model for these scans specifically, this obstacle could be overcome. The text recognition model achieved a character error rate of only 0.56% on held-out test data, corresponding to a word-level accuracy of 96-97%.

The text recognition model was designed to aim for a diplomatic transcription, meaning that the output reflects the text on the page one-on-one, with no standardization. This is notable because eighteenth- and nineteenth-century Danish included a high degree of orthographic variation, which only gradually moved toward homogeneity. No post-OCR correction was applied.

In addition to transcription, newspaper editions were segmented into individual texts. This was tackled as a line-level classification task: for each line, a model predicted whether the text should be split before that line. The prediction was done based on a series of features including the output of three `SetFit` models that each predicted whether a line represented a header, the first line of a text, or the last line of a text. For the final prediction, information on neighboring lines was also included. Additionally, the final model (a `Random Forest` classifier) was designed to take grammatical features such as capitalization and full stops at the end of lines, the presence of frequently recurring place names, and the number of characters into consideration. Because this classification happens at line level, the workflow is less prone to error when dealing with short texts, while long texts will risk being split into standalone paragraphs. It also deals better with texts in recurring formats. Yet, while far from perfect, this approach represents a relatively light-weight and repeatable way to get to the base unit that would interest most scholars, the individual article, rather than columns, pages, or editions, which might all hold many individual texts.

The resulting digital corpus showed substantial improvement. Despite these measures, some variation in accuracy remained. For instance, errors in layout recognition sometimes tripped up line- or word order. Further, a decisive factor for the text recognition itself proved to be the condition of the paper: in damaged areas, recognition quality drops

[5]The text recognition model 'Danish Newspapers 1750-1850' is available to try out on Transkribus.

noticeably, sometimes beyond what even human readers can decipher. To account for this, each text was assigned a predicted word accuracy score, calculated by matching its tokens against a historical dictionary of Danish spelling variants (compiled from eighteenth- and nineteenth-century sources). Across the corpus, the median score is 95.8%, with interquartile values of 92.9% and 98.1%. While coarse, this metric provides a useful per-text indicator of model performance, enabling filtering by accuracy in downstream research.

4.2. Benchmark tasks

Before enriching the corpus with document embeddings, we performed three benchmark tasks to assess their quality, testing six embedding models in total. Two of these models are fine-tuned on Danish historical texts: `MeMo-BERT-03` is fine-tuned on Danish and Norwegian novels from the period 1870–1900, and `Old_News_Segmentation_SBERT_V0.1`[6] is used for segmentation of the newspaper corpus used in this study. In addition to that, we tested the `multilingual-e5-large`[7] model. All three models have shown potential in earlier studies on Danish historical texts (Al-Laith et al., 2024; Feldkamp et al., 2024; Lassche et al., 2025). We furthermore included three state-of-the-art models (`jina-embeddings-v3`[8], `bge-m3` and `embeddinggemma-300m`[9]) that take a maximum input length of 8,194 tokens, allowing for most of the newspaper articles to be represented in one embedding.[10] For the other models, which take a maximum input length of 512 to 514 tokens, we split articles into chunks of up to 514 tokens. A mean embedding was computed by averaging across the resulting chunk embeddings.[11]

4.2.1. Article category classification

In this first benchmarking experiment, we perform an article classification task. We created a gold sample by manually annotating articles for their article category. The three article categories that are present in the sample are *National news*, *International news*, and *Advertisement*. Our sample contained 500 articles from each category, which are randomly taken from 11 newspapers in our corpus of which we know that all three categories are commonly represented.[12] We trained a `LogisticRegression` model in 50 iterations, using a 80/20 train-test split.[13] We started with a baseline experiment, in which we used TF/IDF representations as features. We used the default `ngram_range` which includes unigrams only, using a `max_features` of 10,000. We assume this to be a rather simplistic representation of the article text. Afterwards, we performed experiments with the semantic embeddings of each model as input features. The average performance metrics of this classification task are presented in Table 2.

The 'clean environment' in the above described task, however, does not fully reflect how the articles in the sample look like. In reality, there are articles that do not belong to any of the three categories described above, but can be considered trash or paratext. Those can be headers of the newspaper or headers of subcategories. Because these are often set in ornamental types they tend to contain more OCR-errors than running prose.[14] To test whether it is possible to detect these miscellaneous articles in a separate category, we performed another classification task, this time adding them as the category *Miscellaneous*. Because this category entails a very small part of the corpus, we did not balance our sample, but worked instead with a sample of 500 articles for the categories *National news*, *International news* and *Advertisement*, and 180 articles from the category *Miscellaneous*. Here

[6] For ease of reference, we refer to this model as `Old_News` throughout the remainder of this paper.

[7] For ease of reference, we refer to this model as `e5` throughout the remainder of this paper.

[8] For ease of reference, we refer to this model as `jina` throughout the remainder of this paper.

[9] For ease of reference, we refer to this model as `gemma` throughout the remainder of this paper.

[10] The three models were chosen for their strong MTEB performance, manageable size (<1B parameters), and non-instruction-tuned nature.

[11] Note that the SentenceTransformer version has been shown to affect `embeddinggemma-300m` embeddings (see https://huggingface.co/google/embeddinggemma-300m/discussions/8). For extracting embeddings, we used the most recent SentenceTransformers version, i.e., v5.1.

[12] These include the following titles: Aalborg Stiftstidende, Aalborgs Stifts Adresseavis, Aarhuus Stifts-Tidende, Berlingske Tidende, Den Nord-Cimbriske Tilskuer, Den Vest-Sjællandske Avis, Efterretninger fra Adresse-Contoiret i Bergen, Jyske Efterretninger, Københavns Adresseavis, Lolland-Falsters Stifts-Tidende, Norske Intelligenssedler, Odense Adresse-Contoirs Efterretninger, Ribe Stifts Adresseaviser, Tronhiems Adresse-Contoirs Efterretninger, and Viborger Samler.

[13] We used the `scikit-learn` package in Python (Pedregosa et al., 2011).

[14] Examples of these include 'Den Vest-Siællandske Avis aller Ugeblad. Slagelse. den for landske Historie. uden, og inden før vigtigste Nyeheder. Dagens Redigeret af Pastor og Ridder Bastholm. Trykt og forlagt af Peter Magnus, 18de Aar. Løverdagen, den 5. Januar 1823. Med Kongl. allernaadigst Tilladelse, at forsendes med Brevposten i Danmark og Hertugdømmene' and shorter fragments like 'Fædrelandet' or 'Frankerig'.

we also trained a `LogisticRegression` model in 50 iterations, using a 80/20 train-test split, with the semantic embeddings of each model as features, using TF/IDF representations as baseline. The average performance metrics of this four-class classification task are presented in Table 3.

4.2.2. Fiction/non-fiction classification

In a second benchmark task, we evaluated the performance of the six embedding models on a fiction/non-fiction classification task, focusing on serialized fiction fragments published in newspapers. We annotated a gold sample of 1,871 articles, labeled either *fiction* ($n = 821$) or *non-fiction* ($n = 950$).[15] For classification, the majority class was downsampled to match the size of the minority class. In the final classification, 1,642 datapoints were used (821 per class).

Some pieces, especially in the fiction (a genre often printed in longer runs of text than the rest of the articles), were split across multiple article fragments. Moreover, some fiction pieces are feuilletons, i.e., parts of a series that run across newspaper editions. We handled these patterns in our data as follows: we assigned an ID to each serialized piece or fragment of the same piece. To ensure consistent grouping, we handled missing or incomplete IDs by assigning them dummy values. For evaluation, we used a `StratifiedGroupKFold` cross-validation procedure, which preserves both the class balance and the integrity of the ID groups across training and test splits. This ensured that parts of a series or fragments of the same piece were not split across the train and test sets, and the overall balance between the labels (fiction/non-fiction) was preserved in each fold. We performed 5-fold cross-validation, and in ein which, for d, a `LogisticRegression` classifier was trained on the training set and evaluated on the held-out fold. Performance metrics (precision, recall, F1-score) were computed per class and then averaged across folds. This procedure was applied to both TF/IDF features[16] and the embeddings from all six models, after removing a small number of invalid embeddings (5 data pointsrect comparison of feature representations while controlling for group-level dependencies. The average performance metrics are presented in Table 4.

[15] Two annotators with a background in literary studies labeled the articles, consulting the original newspaper scans. In most cases, headings made the distinction clear (e.g., 'Anecdote' marked short, humorous, and newsworthy renderings, rather than fiction). We retained subcategories that preserve the complexity of the task, such as 'biography' and 'travelogue', which occur under both fiction and non-fiction labels.

[16] Again, with 10,000 max features and unigrams only.

4.2.3. Feuilleton clustering

In the third benchmark task, we clustered the feuilleton articles. We used both fiction and non-fiction articles, and retained only those that had an ID, i.e., were continued in the same newspaper or in a subsequent edition. We applied K-Means clustering across the different embedding models, setting the number of clusters equal to the number of unique series IDs (163). Clustering performance was quantified using the Adjusted Rand Index and V-measure. This procedure allowed us to evaluate how effectively different embedding models capture the natural groupings of pieces and feuilletons (continuous series) and to compare their performance across the corpus systematically. The results of this task are presented in Table 5.

5. Discussion

5.1. Benchmark tasks performances

The results of the first benchmark task show that performances are very close to each other, and that the TF/IDF baseline is only slightly outperformed by the embedding models, except for `gemma`. The best results in article category classification are achieved by the historical Danish model `Old_News`, which surpasses all multilingual models – both the shorter-input model (max. 514 tokens) and those with longer input capacity (8,194 tokens). This holds for both the three-class task (excluding *Miscellaneous*) and the four-class task (including it). Adding the *Miscellaneous* class does in each case decrease the performance of the models, but in the case of the `Old_News`, the `MeMo-BERT-03`, and the `bge-m3` model, this effect is the smallest.

The results of the second benchmark task, involving a fiction/non-fiction classification task, partly align with the results from the first benchmark task, although they differ from it in other aspects. In the feuilleton classification task, it is again the `Old_News` model that outperforms all other models. This time, the `e5` model is a close second. The models `jina` and `bge-m3` obtain in the fiction category only marginally better F1-scores than the TF/IDF baseline – for the non-fiction category, the baseline is not reached. The `gemma` model is clearly underperforming with regards to the baseline, a trend similar with the one observed in the first benchmark task. The results especially show how solid of a predictor the TF/IDF representations are in both classification tasks.

The solid performance of the `Old_News` model aligns with a broader trend towards small, specialized, and efficient models. Such models are particularly attractive when they outperform large multilingual ones (like `jina`), as they can be hosted locally at lower financial and environmental cost.

Model	Precision			Recall			F1-score		
	ads.	int.	nat.	ads.	int.	nat.	ads.	int.	nat.
TF/IDF	0.937	0.885	0.853	0.926	0.952	0.878	0.931	0.917	0.865
MeMo-BERT-03	0.939	0.951	0.902	0.937	0.956	0.897	0.937	0.954	0.899
Old_News	0.969	0.972	0.938	0.958	0.974	0.944	0.963	0.973	0.941
e5	0.945	0.926	0.890	0.927	0.957	0.874	0.935	0.941	0.881
jina	0.926	0.936	0.870	0.915	0.939	0.875	0.920	0.937	0.872
bge-m3	0.949	0.951	0.884	0.917	0.956	0.907	0.932	0.953	0.895
gemma	0.832	0.845	0.775	0.821	0.879	0.750	0.825	0.861	0.762

Table 2: Performance metrics for the three-class classification task using `LogisticRegression` models with different embeddings as features and TF/IDF representations as baseline. The three classes are *Advertisement* (*ads.*), *International news* (*int.*) and *National news* (*nat.*). The dataset is down-sampled to have balanced classes (500 articles per class). 20% was used for testing. Performances are averaged over 50 iterations.

Model	Precision				Recall				F1-score			
	ads.	int.	nat.	misc.	ads.	int.	nat.	misc.	ads.	int.	nat.	misc.
TF/IDF	0.937	0.885	0.853	0.996	0.926	0.952	0.878	0.663	0.931	0.917	0.865	0.792
MeMo-BERT-03	0.933	0.950	0.889	0.928	0.931	0.955	0.896	0.884	0.931	0.952	0.892	0.903
Old_News	0.958	0.971	0.930	0.933	0.955	0.974	0.936	0.901	0.956	0.972	0.932	0.915
e5	0.915	0.907	0.819	0.968	0.922	0.960	0.873	0.508	0.918	0.932	0.845	0.660
jina	0.902	0.915	0.823	0.957	0.906	0.938	0.875	0.641	0.904	0.926	0.848	0.765
bge-m3	0.940	0.940	0.858	0.960	0.906	0.960	0.915	0.787	0.922	0.950	0.885	0.862
gemma	0.810	0.837	0.758	0.840	0.812	0.880	0.757	0.675	0.811	0.857	0.757	0.746

Table 3: Performance metrics for the four-class classification task using `LogisticRegression` models with different embeddings as features and TF/IDf representations as baseline. The four classes are *Advertisement* (*ads.*), *International news* (*int.*), *National news* (*nat.*) and *Miscellaneous* (*misc.*). The dataset is down-sampled to have balanced classes (500 articles per class, except for the *Miscellaneous* class). 20% was used for testing. Performances are averaged over 50 iterations.

However, the other specialized model tested here – `MeMo-BERT-03`, trained on nineteenth-century Danish novels – falls short of this promise. Its performance is only on par with the large multilingual `bge-m3` model. This result likely reflects the limited training data (about 850 novels), which restricts the model's generalizability beyond its narrow domain.

Surprisingly, the feuilleton clustering task (the third benchmark task) shows us results that deviate from the earlier tasks. While `Old_News` performed best in the first two benchmark tasks, `jina` outperforms most models in this clustering task, with `bge-m3` as a runner up (although with less than half of the ARI score of `jina`). This may partly reflect the larger **maximum token input** of `jina` and `bge-m3` (8,192 vs. 512 for `Old_News`), as a larger context window allows the models to integrate longer passages and capture cross-sentence relations, which helps cluster fragments of articles or pieces of a series together. However, consider that `gemma` also has relatively large max tokens (2,048), but is at the bottom of performance in every task.

Embedding **dimensionality** may also play a role: models with 1,024 dimensions (`jina`, `bge-m3`, `e5`) may encode richer semantic detail than those with 768 dimensions (`MeMo-BERT-03`, `Old_News`, `gemma`), aiding the fine-grained grouping of fragments. Notably, `jina`, `bge-m3`, and `e5` are all derived from XLM-RoBERTa-large, which are the three models with the highest ARI scores. This suggests that a strong **pretrained backbone** may provide rich, generalizable representations that support this type of nuanced clustering.

The above observations suggest that context length, embedding dimensionality, and backbone architecture interact in complex ways with the training data and objectives of the six models evaluated. The outcome of this interaction varies across models, meaning that individual advantages do not necessarily translate into consistent overall gains. For example, `e5` and `Old_News` have similarly low maximum token lengths (514/512), yet `e5`'s XLM-RoBERTa-large backbone with higher dimensionality (1,024 vs. 768) achieves comparable performance to the domain-trained `Old_News` (slightly lower V-score, slightly higher ARI). The best model for this task, `jina`, combines a large context window (8,192), high-dimensional embeddings (1,024), and an XLM-RoBERTa-large backbone (with LoRA modifications), likely explaining its consistently strong clustering performance for this task.

Backbone architecture and embedding dimensionality likely contribute to the generally strong performance of `bge-m3` and `e5` in the first benchmark task (article category classification), where they are runner-ups to the domain-specific models, and in the second task (fiction/non-fiction classification), where `e5` performs on par with `Old_News`.

Model	Precision		Recall		F1-score		Accuracy
	fiction	non-fiction	fiction	non-fiction	fiction	non-fiction	
TF/IDF	0.889	0.837	0.820	0.904	0.853	0.869	0.863
MeMo-BERT-03	0.8749	0.882	0.879	0.880	0.878	0.880	0.880
Old_News	0.890	0.883	0.881	0.891	0.885	0.887	0.886
e5	0.884	0.871	0.867	0.887	0.874	0.878	0.876
jina	0.863	0.868	0.869	0.862	0.865	0.865	0.865
bge-m3	0.865	0.871	0.870	0.866	0.867	0.868	0.867
gemma	0.827	0.817	0.816	0.830	0.821	0.823	0.822

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

Model	ARI	V
TF/IDF	0.078	0.618
MeMo-BERT-03	0.100	0.613
Old_News	0.086	0.659
e5	0.114	0.618
jina	0.262	0.785
bge-m3	0.122	0.676
gemma	0.034	0.055

Table 5: Clustering performance of different embedding models on fragment of the same story or parts of series. We only used the datapoints from the classification task that had a continuation, i.e., 163 series in total. Rows are colored based on their performance (V score).

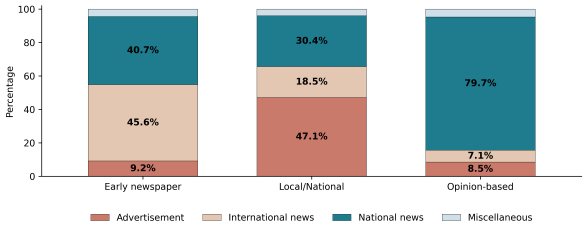

Figure 2: Average category distribution by newspaper groups in the corpus.

Yet these factors do not account for the mediocre performance of `jina`, suggesting that the tasks rely on different aspects of the embeddings – for instance, greater context awareness and semantic detail may be more important for clustering, but not necessarily for classification. This might also explain why `gemma` performs poorly across tasks, as it combines several disadvantages: it is not domain-specific, lacks an XLM-RoBERTa-large backbone, and uses lower-dimensional embeddings than the other multilingual models.

5.2. Final embedding model and its potential

Building on the previous observations about the influence of backbone architecture, dimensionality, and other model specifications, an important takeaway from the three benchmark tasks is that the optimal model choice ultimately depends on the specific analytical goal. Historical data do not necessarily benefit most from historically trained embeddings – instead, corpus, task, and model need to align in both temporal and topical scope. Such alignment does not guarantee that a historically adapted model will perform best on historical material. The third benchmark task, for instance, showed that multilingual models were better suited for clustering, where domain specificity alone proved insufficient.

The two classification tasks further indicated that TF/IDF representations remain a strong baseline and offer a quick, simple, and cost-effective solution for certain applications.[17] Given the purpose of this paper – introducing a corpus enriched with embeddings intended to support further computational analysis – we select the `Old_News` embeddings as the most generally applicable. Its strong results in both classification tasks suggest that this model is the most promising candidate for future research using this corpus.

As a means of illustration, we predict article categories using the classifier from the first benchmark task to verify a simple historical hypothesis: that different types of papers exhibit distinct tendencies to privilege advertisements, national or international news.[18] As shown in Figure 2, the

[17] Note that our use of 10,000 TF–IDF features results in a strong baseline. The dimensionality is substantially higher than that of many embedding models (e.g., 768 dimensions), making it more likely that discriminative lexical cues are captured. This is particularly relevant for tasks such as fiction vs. nonfiction classification, where simple surface features – like personal pronoun frequency – already provide strong signals (Qureshi et al., 2019; Kazmi et al., 2022; Feldkamp et al., 2025).

[18] We used the four-class `LogisticRegression` classifier with the `Old_News` embeddings as features.

three genres outlined in Table 1 indeed diverge in meaningful ways, reflecting their historical functions: early, more elite-oriented newspapers emphasize international news, later local papers, with a more commercial profile, foreground advertisements, while opinion-based papers focus on national issues and commentary.[19]

6. Conclusions and Future Work

In this paper, we presented an enriched dataset of Danish historical newspapers spanning the late seventeenth to the nineteenth century, comprising nearly five million articles annotated with article-level semantic embeddings. To evaluate how well language models represent historical Danish, we conducted three benchmark tasks – article category classification, fiction/non-fiction classification, and feuilleton clustering – that compared the performance of six Danish and multilingual embedding models. Our findings suggest that, beyond domain specificity in temporal and topical scope, architectural factors such as maximum token input length, embedding dimensionality, and backbone design can significantly influence model performance on historical corpora.

Based on these results, we enriched the full dataset with embeddings from `Old_News`, which proved the most reliable and interpretable model across tasks. We demonstrated their potential by testing a hypothesis about subgenres in the corpus. Using the annotated gold sample from the first benchmark task, we predicted article categories across the entire corpus, confirming differences in category distribution among these subgenres. This small demonstration illustrates how the resource can support historically informed, data-driven research on the evolution of Danish print culture and the Danish language more broadly. The article categories enable more targeted investigations of national and international news, for instance by examining the geographical horizons of reporting and the flow of information. In turn, the annotated gold samples from the second and third benchmark tasks provide opportunities for focused analyses of fictional content distributed through the newspapers.

By releasing both the corpus and its embeddings openly, we hope to enable reproducible, cross-disciplinary research on the Danish nineteenth-century mediascape and contribute to broader efforts in computational historical linguistics and digital cultural heritage.

[19]The predicted labels are included in the `predicted_category` column of the dataset, available via https://huggingface.co/datasets/chcaa/eno-embs-old-news.

Limitations

The quality of the **OCR transcriptions** is dependent on the physical qualities of the original sources. Newspapers were printed in *fraktur* on cheap, thin paper, and later microfilmed – in some cases as early as the 1950s – all of which has left its mark. The uneven type, worn paper, and scanning noise are not only technical obstacles but also part of the material history of nineteenth-century print culture.

The analyses presented in this study rely on **article categories** predicted by a classifier with high, but not perfect, performance. As a result, some articles may have been incorrectly assigned to categories such as *National news*. Moreover, the four-class classification task necessarily constrains all articles to fit within one of these predefined categories, even when certain texts – such as fiction – do not naturally belong to any of them.

In addition to our discussion about **model specifications**, other characteristics of the embedding models or the texts, which are not discussed in detail here, may have influenced their performances. Future work could explore how these model-specific factors affect performance across different article categories and time periods.

7. Bibliographical References

Anne Agersnap, Katrine Baunvig, Line Wittoff Schmidt, Rie Schmidt Eriksen, Emil Walther Bønding, Thomas Husted Kirkegaard, and Lea Wierød Borčak. 2025. The Human Touch: Leveraging HITL for Quantitative Close Reading of Historical Corpora. In *DHNB 2025 Book of Abstracts*, pages 11–13, Tartu.

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.

Benedict R. O'G Anderson. 2016. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, revised edition edition. Verso, London New York.

Katrine Frøkjær Baunvig. 2021. Fictional Realities of Modernity: The Fantastic Life of Demi-Goddess Dana in the Emerging Nation State of Denmark. In Sophie Bonding, Lone Kolle Martinsen, and Pierre-Brice Stahl, editors, *Mythology*

and Nation Building: N.F.S. Grundtvig and His Contemporaries. Aarhus University Press.

Katrine Frøkjær Baunvig. 2023. "Each of Our Springs Has Lost Its Miraculous Power": The Range of a Religious Hotspot – A Distant Reading of Lourdes Representations in Denmark 1858–1914. *Numen*, 70(1):43–69.

Molly Brandt Skelbye and Dana Dannélls. 2021. OCR Processing of Swedish Historical Newspapers Using Deep Hybrid CNN–LSTM Networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 190–198, Held Online. INCOMA Ltd.

Jørgen Burchardt. 2024. Source criticism, bias, and representativeness in the digital age: A case study of digitized newspaper archives. *Digital Humanities in the Nordic and Baltic Countries Publications*, 6(1). Number: 1.

Kevin Cohen, Laura Manrique-Gómez, and Ruben Manrique. 2025. Historical Ink: Exploring Large Language Models for Irony Detection in 19th-Century Spanish. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 559–569, Albuquerque, USA. Association for Computational Linguistics.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: The Impresso Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding.

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

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

Claus Friisberg. 2013. Den kildekritiske tvang. *Historisk Tidsskrift*, 100(2).

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

Johan Heinsen and Anders Dyrborg Birkemose. 2023. Efterlyst. Identitet, tvang og mobilitet, 1750-1850. *Temp - tidsskrift for historie*, 14(27):24–53.

Johan Heinsen and Camilla Bøgeskov. 2025. A World in Print: Introducing a Danish-Norwegian corpus of historical newspapers.

Henrik Horstbøll. 1999. *Menigmands medie: det folkelige bogtryk i Danmark 1500-1840 : en kulturhistorisk undersøgelse*. Danish Humanist Texts. Kongelige bibliotek.

H. Jensen, F.G. Laustsen, Denmark. Rigsdagen, and Copenhagen Institut for historie og samfundsøkonomi. 1934. *De danske stænderforsamlingers historie, 1830-1848: del. Stænderforsamlingernes virksomhed og betydning fra 1838 indtil 1848*. De danske stænderforsamlingers historie, 1830-1848. J. H. Schultz.

T.G. Jørgensen. 1956. *Andreas Frederik Krieger: juristen, politikeren, borgeren*. A. Frost-Hansen.

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

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

Thorkild Kjærgaard. 1989. The rise of press and public opinion in eighteenth-century Denmark—Norway. *Scandinavian Journal of History*.

Mika Koistinen, Kimmo Kettunen, and Tuula Pääkkönen. 2017. Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In *Proceedings of the*

21st Nordic Conference on Computational Linguistics, pages 277–283, Gothenburg, Sweden. Association for Computational Linguistics.

Alie Lassche, Pascale Feldkamp, Yuri Bizzoni, Katrine Baunvig, and Kristoffer Nielbo. 2025. Why Novels (Don't) Break Through: Dynamics of Canonicity in the Danish Modern Breakthrough (1870-1900). In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 278–290, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Viktoria Löfgren and Dana Dannélls. 2024. Post-OCR Correction of Digitized Swedish Newspapers with ByT5. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 237–242, St. Julians, Malta. Association for Computational Linguistics.

Devon Mordell. 2019. Critical Questions for Archives as (Big) Data. *Archivaria*, pages 140–161.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Fabian Pedregosa, Fabian Pedregosa, Gael Varoquaux, Gael Varoquaux, Normalesup Org, Alexandre Gramfort, Alexandre Gramfort, Vincent Michel, Vincent Michel, Logilab Fr, Bertrand Thirion, Bertrand Thirion, Olivier Grisel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, Alexandre Tp, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Andrew Pettegree. 2014. *The Invention of News: How the World Came to Know about Itself*. Yale University Press, New Haven; London, England.

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

Teemu Ruokolainen and Kimmo Kettunen. Name the name – named entity recognition in OCRed 19th and early 20th century finnish newspaper and journal collection data. In *DHN 2020. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pages 137–156.

Jette D. Søllinge and Niels Thomsen. 1988. *De danske aviser 1634-1989 (I)*. Dagspressens Fond, Odense.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for Post-OCR Correction of Historical Newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Crina Tudor, Beata Megyesi, and Robert Östling. 2025. Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226, Albuquerque, New Mexico. Association for Computational Linguistics.

Johannes Weber. 2006. Strassburg, 1605: The Origins of the Newspaper in Europe. *German History*, 24(3):387–412.

Appendix A. Detailed corpus statistics

Newspaper	Period	Editions	Articles	Words
Aalborg Stiftstidende	1827–1846	4,519	203,072	1,9026,204
Aalborgs Stifts Adresseavis	1818–1827	1,591	62,540	5,212,682
Aarhuus Stifts-Tidende	1794–1847	8710	365738	30492484
Adresseavis for Børn	1779–1782	206	1,471	504,229
Almuevennen	1842–1844	161	2,020	708,770
Berlingske Tidende	1749–1836	10,946	792,663	113,926,106
Corsaren	1840–1844	211	2,832	866,368
Danske Mercurius	1666–1691	144	2,308	256,423
Den Nord-Cimbriske Tilskuer	1824–1848	3,238	94,609	9,522,907
Den Vest-Sjællandske Avis	1815–1842	3,621	132,439	16,824,428
Efterretninger fra Adresse-Contoiret i Bergen	1765–1814	2,175	86,338	5,740,796
Extraordinaire Maanedlige Relationer	1672–1698	250	7,372	876,360
Extraordinaire Relationer	1720–1748	338	6,887	714,399
Jyllandsposten	1838–1841	250	3,471	933,957
Jyske Efterretninger	1767–1823	4,485	150,227	14,433,761
Kiøbenhavns Extraordinaire Relation	1721–1745	341	2,796	485,171
Kiøbenhavns Maanetlige Postrytter	1731–1748	213	4,172	470,029
Kiøbenhavns Postrytter	1733–1798	1,920	46,860	4,268,815
Københavns Adresseavis	1759–1837	11,425	1,377,081	105,900,429
Lolland-Falsters Stifts-Tidende	1809–1847	4,435	163,765	14,303,036
Norske Intelligenssedler	1763–1814	2,957	175,327	12,426,028
Nye Tidender	1698–1730	54	891	111,246
Nyeste Skilderie af Kiøbenhavn	1803–1831	2,891	73,059	13,124,176
Odense Adresse-Contoirs Efterretninger	1772–1848	12,155	520,007	45,674,584
Politivennen*	1798–1809	-	4,051	1,071,413
Ribe Stifts Adresseaviser	1786–1848	6,132	238,260	20,565,322
Tronhiems Adresse-Contoirs Efterretninger	1767–1799	1,683	33,808	3,012,461
Viborger Samler	1773–1849	8,845	344,020	31,684,297
Total		93,908	4,898,084	473,136,881

Table 6: Descriptive statistics on the newspapers and periodicals in the corpus. *Since the editions of the *Politivennen* were undated, we only provide the total number of articles and words for this periodical.

Appendix B. Model specifications

Model	Max tokens	Dimensions	Layers	Derived from	Source
MeMo-BERT-03	514	768	12	DanskBERT	https://huggingface.co/MiMe-MeMo/MeMo-BERT-03
Old_News_Segmentation_SBERT_V0.1	512	768	12	DA-Bert_Old_News_V1	https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0.1
bge-m3	8,192	1,024	24	XLM-RoBERTa-large	https://huggingface.co/BAAI/bge-m3
embeddinggemma-300m	2,048	768*	24	T5Gemma	https://huggingface.co/google/embeddinggemma-300m
jina-embeddings-v3	8,192	1,024*	24	XLM-RoBERTa-large**	https://huggingface.co/jinaai/jina-embeddings-v3
multilingual-e5-large	514	1,024	24	XLM-RoBERTa-large	https://huggingface.co/intfloat/multilingual-e5-large

Table 7: Full model names, maximum input context length, final embedding dimension size, number of hidden layers, information on which model a given model is derived from, and HuggingFace urls. The order of models is by language (Danish models on top) and alphabetical. *Output embedding dimension size is flexible (smaller is supported). **While jina-embeddings-v3 is based on XLM-RoBERTa-large, it is augmented with LoRA adapters and fine-tuned for embedding tasks. 'Derived from' indicates the pretrained backbone or base model.