The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding

Anonymous ACL submission

Abstract

The evaluation of English text embeddings has transitioned from evaluating on a handful of datasets to broad coverage across many tasks through benchmarks such as MTEB. However, this is not the case for multilingual text embeddings due to a lack of available benchmarks. To address this problem, we introduce the Scandinavian Embedding Benchmark (SEB). SEB is a comprehensive framework that enables text embedding evaluation for Scandinavian languages across 24 tasks, 10 subtasks, and 4 task categories. Building on SEB, we evaluate more than 26 models, uncovering significant performance disparities between public and commercial as well as monolingual and multilingual text embedding models. We open-source SEB[1] and integrate it with MTEB, thus bridging the text embedding evaluation gap for Scandinavian languages.

1 Introduction

Natural language embeddings are used in a diverse range of applications, including clustering (Liu and Xiong, 2011; Angelov, 2020), text mining (Jiang et al., 2015), semantic search (Reimers and Gurevych, 2019a; Muennighoff, 2022) and feature representation (Alayrac et al., 2022). Furthermore, embeddings are crucial in retrieval augmented generation (RAG) systems (Borgeaud et al., 2022), particularly for low- to mid-resource languages and domains. RAG systems enable the enrichment of generative models with the knowledge that might be underrepresented or absent during training. Thus, they can play a role in broadening linguistic and domain coverage.

[1] https://anonymous.4open.science/r/scandinavian-embedding-benchmark-88C0

With the breadth of applications for text embeddings, a proper evaluation of their quality is critical. Recent work has proposed Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023), a benchmark for evaluating the quality of document embeddings for a wide variety of tasks. MTEB improves upon prior benchmarks by addressing the lack of evaluations across tasks. This has led to the widespread adoption of the benchmark for evaluating natural language embeddings.

However, while MTEB substantially improves the evaluation of text embeddings, the benchmark has the following shortcomings:

1. MTEB contains only limited support for evaluating non-English embeddings, especially across a wide range of tasks.

2. Furthermore, MTEB does not include model implementations in the benchmark's code. This makes the results on the leaderboard hard to reproduce[2]. This is especially problematic for prompt-based embedding models (Muennighoff, 2022; Xiao et al., 2023; Su et al., 2022) where the prompt of choice can significantly impact performance.

3. While MTEB has broad coverage across tasks, its domain coverage is still limited, as it primarily includes datasets from academic articles, social media, and web sources.

1.1 Contributions

To mitigate these issues, we present SEB a benchmark for embedding evaluation of the Mainland Scandinavian languages: Danish

[2] This can, for instance, be seen in issues such as https://github.com/embeddings-benchmark/mteb/issues/109

(da), Swedish (sv), and Norwegian (Bokmål (nb) and Nynorsk (nn)) as well as the Danish dialect Bornholmsk (da-bornholm). This initiative is supported by findings from a study by Nielsen (2023), which demonstrates substantial cross-lingual transfer between these languages; this supports collectively benchmarking the Mainland Scandinavian languages to broaden the coverage otherwise limited for these languages. SEB makes the following main contributions; 1 it greatly expands the evaluation of embedding for Scandinavian to multiple tasks (see Table 1) as well as across a wide range of domains (see Table 2); 2 SEB implements a model registry that allows for the easy addition of new models as well as documentation of the exact implementation of existing models evaluated in the benchmark. Lastly, 3 SEB expands and extends MTEB by porting all tasks, allowing for the expansion of MTEB to a fully-fledged multilingual benchmark for embeddings. Using SEB we evaluate 26 representative models and APIs within this work and present additional models in an interactive online dashboard.[3]

2 Related Work

2.1 Benchmarks

Benchmarks are important tools for model development that enable the assessment of significant performance improvements. Prior benchmarks for evaluating text embeddings focused on specific embedding qualities; BEIR (Thakur et al., 2021) and MIRACL (Zhang et al., 2023) assessed embedding efficacy in information retrieval across diverse domains or languages, while SentEval (Conneau and Kiela, 2018) integrated various SemEval datasets for sentence encoding evaluation using semantic text similarity (STS) tasks. MTEB (Muennighoff et al., 2023) amalgamated and expanded these methodologies to cover eight different tasks. While MTEB includes more than 112 languages, most of this linguistic variation originates from only a handful of tasks, notably bitext mining (Tatoeba Project Contributors, 2023) or translated datasets (FitzGerald et al., 2022). Scandinavian languages are only represented in two datasets for intent and scenario classification (FitzGerald et al., 2022), both of which are translations. Thus, the benchmark contains no naturally occurring text for either of these languages.

While benchmarks for Scandinavian languages have been developed, most – akin to (Super)GLUE (Wang et al., 2018, 2019) – seek to evaluate the performance of multiple natural language understanding tasks. These include monolingual benchmarks such as the Swedish superlim (Berdicevskis et al., 2023), the Norwegian NorBench (Samuel et al., 2023), or cross-lingual benchmarks such as ScandEval (Nielsen, 2023). While these benchmarks are instrumental for developing Scandinavian models, none focus on evaluating text embeddings for, e.g., retrieval or clustering.

2.2 Text Embeddings

Over time, the development of dense text embedding models has evolved from focusing on individual words (Mikolov et al., 2013; Pennington et al., 2014) to encompass entire sentences (Conneau et al., 2017; Ni et al., 2021), and currently extends to processing multiple sentences in a wide range of tasks (Xiao et al., 2023; Su et al., 2022). As is common in natural language processing (Xue et al., 2020), English-centric models have led this development, followed by multilingual models with only a short delay. While word-specific and sentence multilingual embedding models already exist (Artetxe and Schwenk, 2019), multitask embedding models are just beginning to emerge (Chen et al., 2024; Wang et al., 2022). However, their progress is hindered by the lack of comprehensive evaluation in multilingual tasks. This evaluation gap hinders progress in the field, preventing us from effectively evaluating model improvements. Our work aims to address this problem to enable further progress and proliferation of multilingual text embedding.

3 The Benchmark

3.1 Design and Curation Rationale

SEB seeks to provide an estimate of the quality of embedding for Scandinavian languages and multilingual use cases. To do so, we focus on **a) Coverage:** The benchmark should cover a wide variety of tasks spanning distinctly different domains, usages, and embedding tasks;

[3] Anonymized

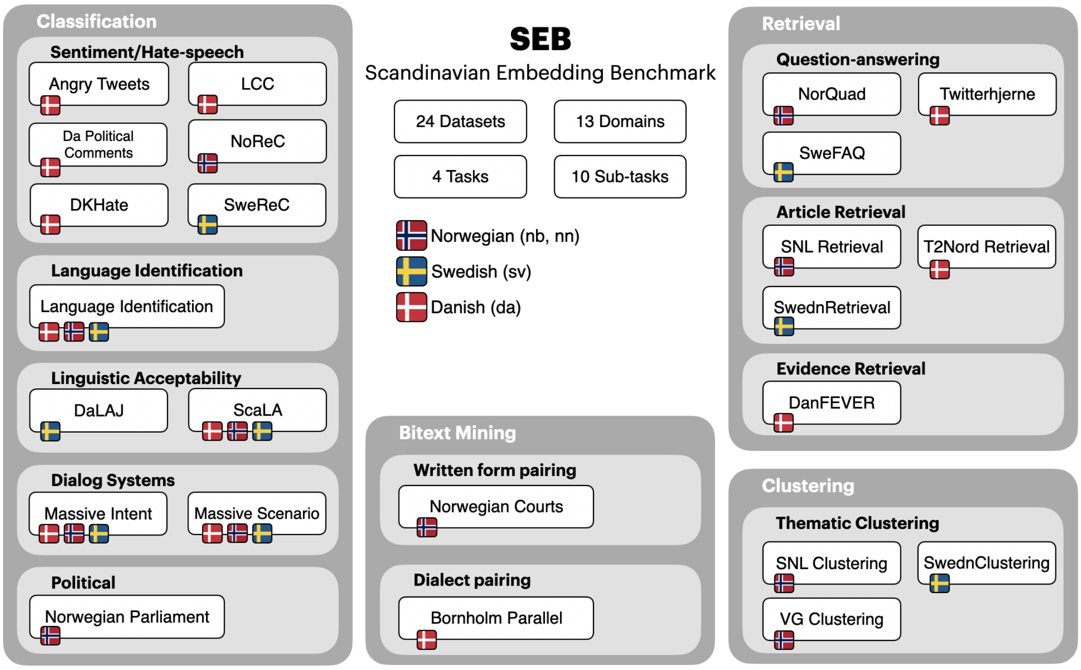

Figure 1: An overview of the tasks and datasets in SEB. Flags denote the languages of the datasets.

SEB compromises 24 datasets spanning at least 12 domains across nine different tasks with broad coverage for each language.

b) Cultural integrity and model equity: Recent studies (Berdicevskis et al., 2023; Nielsen, 2023; Muennighoff et al., 2023) have increasingly adopted the strategy of leveraging translated English datasets as a means to evaluate the performance of models in low-resource language contexts. However, we avoid adding such translations, aiming to represent Scandinavian contexts accurately and mitigate the risk of artificially inflating multilingual model capabilities. This decision stems from the recognition that multilingual models, often trained on parallel or translated data (Reimers and Gurevych, 2020), may exhibit inflated performance when evaluated on similar translated tasks — a hypothesis that, while plausible, remains to be conclusively shown. We choose to keep the existing translated datasets from MTEB within SEB to maintain compatibility.

c) Cross-lingual generalization: Given the limited availability of datasets for the Scandinavian languages, we rely on the high degree of cross-lingual transfer (Nielsen, 2023) to estimate model performance more accurately. This approach capitalizes on intrinsic linguistic similarities and shared cultural contexts to bridge data gaps.

d) Reproducibility and Accessibility: SEB expands upon the reproducibility of MTEB by including a model registry for all evaluated models to ensure the exact method (e.g., model prompts) for obtaining the results is known. Furthermore, to ensure that the benchmark is as widely accessible as possible, we have limited the size of most datasets to a maximum of 2048 examples. For most models, this allows running the benchmark on a consumer-grade laptop while ensuring proper performance estimation. The benchmark also implements a public cache, allowing users to experiment without needing to rerun models run by others.

In addition to these criteria, SEB follows the desiderata outlined by Muennighoff et al. (2023), allowing for easy extension of the benchmark and providing a simple API and command-line interface making it easy to benchmark models that are not part of SEB by default.

3.2 Datasets

We present an overview of the tasks in SEB in Figure 1. Additionally, we have created an overview of the datasets in Table 6, including dataset statistics and a short description of each dataset. subsection A.4 described the method of evaluation, and subsection A.5

described the formalization of the specific datasets to the task. SEB seeks to cover a large variety of domains and task types, greatly expanding upon what was previously available for non-English languages within MTEB (see Table 2 and 1). To allow for the exploration, we add an embedding map of samples from the dataset in subsection A.3, where it is clearly seen that the datasets occupy different clusters. Similarly, Figure 2 reveals distinctly different clusters of datasets, e.g., the high similarity between SNL Retrieval and NorQuad as both are constructed from encyclopedic sources while distinct datasets such as SweFAQ (Berdicevskis et al., 2023), covering FAQ related to the public sector.

	Language			
Domain	da	nb	nn	sv
Academic	(+)			
Bible				
Blog				
Fiction	+	+	+	+
Government	+	+	+	+
Legal	(+)	+	+	
Medical				
News	+	+		+
Non-fiction	+	+		+
Poetry	(+)			
Reviews		+		
Social	+			+
Spoken	✓	✓		✓
Wiki	+	+	+	+
Web	+			+

Table 2: Domain coverage on SEB for Mainland Scandinavian languages. The green plus (+) indicates newly added domains in SEB, while black checks (✓) indicate domains covered in MTEB for Scandinavian Languages. The parenthesis is due to the LCC (Nielsen, 2016) containing the domains, but only to a limited extent. The domains follow the categorization of the Universal Dependencies (Nivre et al., 2017).

	Language			
Task	da	nb	nn	sv
Retrieval				
Question answering	+	+		+
Article retrieval	+	+		+
Bitext Mining				
Dialect pairing	+	+	+	+
Classification				
Political		+	+	+
Language Identification	+	+	+	+
Linguistic Acceptability	+	+	+	+
Sentiment/Hate Speech	+	+		+
Dialog Systems	✓	✓	✓	✓
Clustering				
Thematic Clustering	+	+		+

Table 1: Task coverage across the Scandinavian languages within SEB. The green plus (+) denote newly added tasks, while black checkmarks (✓) denote tasks previously in MTEB.

4 Results

4.1 Models

For our benchmarked models, we have chosen a series of representative models seeking to cover a range of model architectures, model sizes, and commercial APIs, as well as models claiming state-of-the-art results on various embedding tasks. In addition, the online dashboard includes additional models not represented here. We group the models into self-supervised and supervised methods.

Self-supervised methods:

Encoders such as BERT models (Devlin et al., 2019) including monolingual or Scandinavian models trained for Danish (Enevoldsen et al., 2023), Norwegian (Kummervold et al., 2021) and Swedish (Rekathati, 2021) as well as the multilingual model XLM-R (Conneau et al., 2020). We also include a SimCSE (Gao et al., 2021) version of the dfm-encoder-large to indicate the potential performance gain by self-supervised pre-training. This model is trained on sentences extracted from the Danish Gigaword (Strømberg-Derczynski et al., 2021) using default parameters[5].

As a candidate for **Static Word Vectors**, we include four fastText (Joulin et al., 2016, 2017; Bojanowski et al., 2017) models for Danish, Swedish, and Norwegian Bokmål and Nynorsk respectively.

Supervised Methods:

For **encoders**, we benchmark LaBSE (Feng et al., 2022), which is based on BERT but further pre-trained on a parallel corpus. Further, we evaluate the multilingual MiniLM models

[5]For exact specification see the model card; `anonymized`

	Avg.	Task-Type Bitext	Class.	Clust.	Retr.	Language da	nb	nn	sv
Num. Datasets (→)	24	2	12	3	7	12	11	3	9
Self-Supervised Models									
dfm-encoder-large	41.4	46.8	56.5	26.9	20.1	47.7	47.4	72.5	43.7
+ SimCSE	**46.6**	50.9	58.4	26.9	**33.7**	**52.2**	51.3	74.3	42.0
xlm-roberta-large	35.3	19.1	54.6	28.1	10.0	39.6	41.3	58.0	44.5
nb-bert-large	46.0	47.3	**59.3**	**35.7**	27.3	46.8	**57.2**	**80.4**	**50.2**
nb-bert-base	42.1	**51.0**	57.0	31.8	18.4	43.6	53.0	79.2	47.7
bert-base-swedish	35.2	39.1	49.7	26.2	13.2	34.0	41.1	62.2	43.6
fasttext-cc-da	37.3	42.4	48.8	21.8	22.7	39.0	43.2	66.4	38.7
fasttext-cc-nn	35.8	47.6	46.2	22.1	20.4	34.6	43.9	69.1	37.1
fasttext-cc-nb	37.5	43.2	48.7	24.2	22.2	37.5	45.6	67.7	38.9
fasttext-cc-sv	36.0	43.3	47.3	22.0	20.4	34.9	41.3	63.4	40.6
Supervised Models									
multilingual-MiniLM-L12	50.0	51.0	53.7	31.7	51.1	49.9	52.7	58.3	50.3
multilingual-mpnet-base	53.2	52.7	56.5	32.7	56.5	53.0	55.8	59.6	53.3
labSE	50.5	69.1	53.6	29.0	48.9	50.9	52.9	59.4	48.7
sentence-bert-swedish	46.6	43.3	51.0	35.6	44.6	43.2	48.2	62.7	54.7
e5-mistral-7b-instruct	60.4	**70.8**	61.7	35.7	66.0	**61.7**	62.9	68.8	60.4
multilingual-e5-large	**60.7**	60.1	**62.5**	34.2	**69.1**	61.1	**63.1**	**73.9**	**62.8**
multilingual-e5-base	57.9	61.4	60.1	34.0	63.5	58.6	60.9	72.0	58.5
multilingual-e5-small	56.4	61.6	58.1	**36.9**	60.3	56.5	58.9	69.5	57.1
translate-e5-large	47.7	50.7	54.7	27.3	43.4	49.0	50.1	59.2	59.2
sonar-dan	43.4	70.5	53.5	19.6	28.6	48.3	46.0	63.7	42.9
sonar-nob	41.5	63.2	52.9	18.5	25.6	45.2	45.9	64.7	42.4
sonar-nno	41.5	65.5	52.8	17.3	25.7	45.5	45.1	63.2	42.6
sonar-swe	42.8	70.7	52.9	19.4	27.6	47.1	45.4	63.1	42.9
Embedding APIs									
text-embedding-3-large	**65.0**	**68.8**	63.5	38.7	**77.9**	**63.7**	**69.0**	**74.7**	**65.5**
text-embedding-3-small	61.0	66.7	59.7	38.3	71.3	59.7	64.7	70.2	60.4
embed-multilingual-v3.0	64.1	64.2	**63.6**	**40.2**	75.2	62.6	68.5	74.1	64.3

Table 3: Performance across task-type categories and languages in SEB. The best score in each model category is highlighted in bold. Additional model evaluation can be found on the public Dashboard[4].

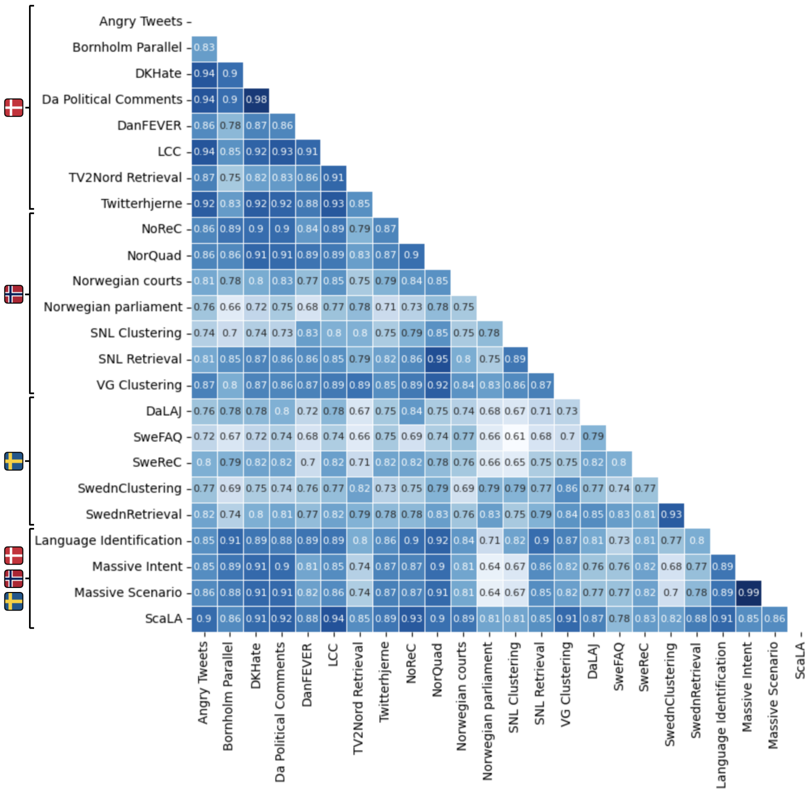

Figure 2: Dataset similarity between the datasets included within SEB. Embeddings are obtained by applying the embed-multilingual-v3.0 on 100 randomly sampled documents. Similarity is computed using cosine similarity.

and MPNet models (Reimers and Gurevych, 2019b; Song et al., 2020; Wang et al., 2021), which are trained on diverse datasets. We also include the SONAR models (Duquenne et al., 2023) as they claim improved performance over LabSE. In addition, we include the Swedish sentence transformers (Rekathati, 2021) trained with knowledge distillation from an English model (Reimers and Gurevych, 2020).
Because the development of Scandinavian **decoders** is only in its early stages (Enevoldsen et al., 2023; Ekgren et al., 2022), we utilize the e5-mistral model (Wang et al., 2022, 2023) as it presents a competitive model in the category.
Commercial embedding APIs: We additionally include the embedding APIs of Cohere [6] and OpenAI [7] to compare openly available models with commercial solutions.
Lastly, we add **Translate and embed** as a baseline model for comparing naïvely translating to English and then embedding with high-quality English models. To allow for comparison with multilingual models, we include both the large English e5 model and all sizes of its multilingual variants (Wang et al., 2022). We use the multilingual M2M100 model (Fan et al., 2020) for the translation. For translation, we assume the language is known. This avoids accumulating errors due to language detection, and in many applications, the language would be known. We assume Danish as the origin for tasks requiring multiple languages, such as bitext mining.

4.2 Analysis

In Table 3, we see that the best-performing model is either of the commercial APIs of OpenAI and Cohere followed by the publicly available multilingual e5 model series (Wang et al., 2022). This stands in contrast to developments observed from ScandEval (Nielsen, 2023), where notably smaller monolingual or Scandinavian models have proven to be competitive, often surpassing significantly larger multilingual models. Similar to MTEB (Muennighoff et al., 2023), we find a pronounced

[6]`https://txt.cohere.com/introducing-embed-v3/`
[7]`https://openai.com/blog/new-embedding-models-and-api-updates`

performance between self-supervised methods and their supervised counterparts, although we see that notable gains can be obtained from unsupervised pre-training (Gao et al., 2021). In general, however, utilizing unsupervised contrastive pretraining pales in comparison to popular multilingual models of smaller size.

In Table 5, we see the performance across domains. Generally, we see that model rankings remain relatively stable across these domains, with the e5 models (Wang et al., 2022) and the commercial APIs taking a consistent lead. However, we also see that in domains such as the legal domain, spoken language, and fiction, we see the e5-mistral-7b-instruct outcompeting commercial solutions.

If the examine individual subtask (see subsection A.7) Pretrained encoders perform surprisingly well on language acceptability and language detection tasks. This is likely due to a trade-off between semantics and syntax. Self-supervised training on natural language will likely assign significance to syntactic nuances, while models trained on semantic tasks ignore some syntactical errors favoring semantics.

Performance across task-types: Models that have been contrastively trained on sentence pairs or finetuned for a set of common tasks typically outperform pre-trained models, especially in retrieval contexts, while LaBSE (Feng et al., 2022) and the SONAR models (Duquenne et al., 2023), which has been designed for bitext-mining purposes, excels at the task.

The largest gap between commercial and public models is in retrieval, where performance drops more than eight points. While notable improvements have been achieved in publicly available embedding models for English retrieval tasks (Wang et al., 2023), similar results are yet to be achieved in multilingual contexts. Bitext mining is the only category in which open solutions outperform commercial solutions.

Translate then embed: When comparing the 'translate-then-embed' model against the multilingual e5 models, we see that in almost all cases, the multilingual models perform better even when comparing across size categories. While performance could likely be improved by utilizing state-of-the-art embedding and translation models, we see few benefits to this approach due to increased computational costs, model complexity, and competitive approaches for knowledge distillation across languages (Reimers and Gurevych, 2020).

4.3 Efficiency

We examine the trade-offs between performance and speed in Figure 3. Speed was benchmarked on Dell PowerEdge C6420 Intel(R) Xeon(R) Gold 6130 CPUs with 32 cores/CPU. We see the following categories of relevance;

Highest Throughput FastText models offer the highest throughput while maintaining an average performance exceeding to that of the multilingual XLM-R (Conneau et al., 2020).

Maximum Performance Achieving optimal performance is possible with the multilingual-e5-large or the e5-mistral-7b-instruct, which rivals the smaller commercial embedding APIs.

Balanced Performance: The best balance between performance, throughput, and embedding size is seen in the multilingual e5 models series, which performs competitively on the benchmark. The multilingual-mpnet-base also offers a competitive balance between throughput and performance, despite its larger embedding size.

4.4 Limitations and Future Perspectives

Domain Coverage: Despite the advancements introduced by SEB, the benchmark could further benefit from encompassing domains crucial to the welfare states of Scandinavia, such as legal, governmental, and medical fields, which are currently only partly covered or unaddressed. Current tasks predominantly feature non-fiction literature, such as encyclopedias and news, yet the rising interest in digital humanities (Su et al., 2020) suggests the inclusion of fiction, poetry, historical texts, and religious documents in future updates could be valuable. Additionally, the benchmark currently lacks some task categories, such as pair classification and document reranking.

Future Directions: While this work announces the release of SEB, we plan to continually expand upon the benchmark. As this work continues to develop, we invite researchers to join us in expanding the evaluation of embedding models across a broad range of languages.

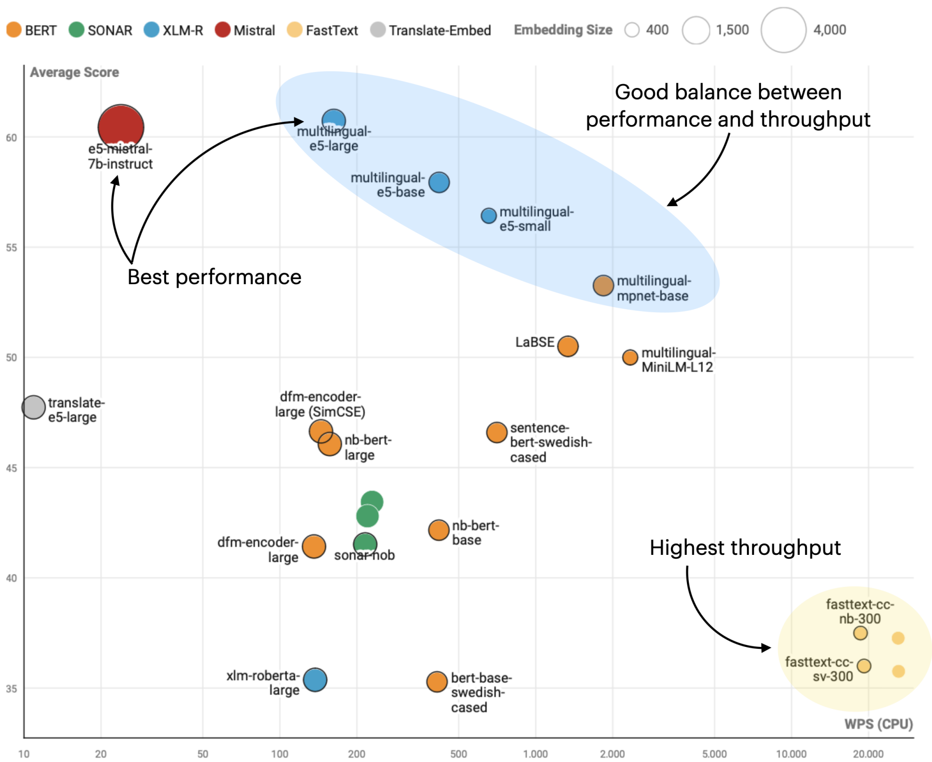

Figure 3: Performance and speed of embeddings models. The size of the circles denotes the embedding size, and color denotes the model type. Note that commercial APIs are not included. WPS stands for words per second. We avoid highlighting all models to increase readability.

5 Conclusion

In this work, we introduced SEB, a framework that addresses the evaluation gap for the mainland Scandinavian languages. SEB encompasses 24 tasks covering ten subtasks in four task categories and spanning mainland Scandinavian languages.

We evaluate more than 50 models on SEB and show that there is still a notable gap in performance between publicly available text embedding models and their commercial counterparts, especially in retrieval contexts, as well as between monolingual and multilingual models. These findings highlight critical areas for future advancements. By open-sourcing SEB and integrating it with MTEB, we aim to encourage the development of robust Scandinavian and multilingual embedding models, inviting the research community to contribute to this evolving landscape.

Acknowledgements

Anonymized (includes acknowledgments of compute resources)

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Dimitar Angelov. 2020. Top2vec: Distributed representations of topics. *ArXiv*, abs/2008.09470.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Leon Derczynski and Alex Speed Kjeldsen. 2019. Bornholmsk natural language processing: Resources and tools. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 338–344, Turku, Finland. Linköping University Electronic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv e-prints*.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

Kenneth Enevoldsen, Lasse Hansen, Dan S Nielsen, Rasmus AF Egebæk, Søren V Holm, Martin C Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B Jørgensen, Malte Højmark-Bertelsen, et al. 2023. Danish foundation models. *arXiv preprint arXiv:2311.07264*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

René Haas and Leon Derczynski. 2021. Discriminating between similar Nordic languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75, Kiyv, Ukraine. Association for Computational Linguistics.

Søren Vejlgaard Holm. 2024. Are gllms danoliterate? benchmarking generative nlp in danish.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid.

2023. NorQuAD: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.

Zhenchao Jiang, Lishuang Li, Degen Huang, and Liuke Jin. 2015. Training word embeddings for deep learning in biomedical text mining tasks. In *2015 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 625–628. IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

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.

Fasheng Liu and Lu Xiong. 2011. Survey on text clustering algorithm-research present situation of text clustering algorithm. In *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, pages 196–199. IEEE.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

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.

Jørgen Johnsen Navjord and Jon-Mikkel Ryen Korsvik. 2023. Beyond extractive: advancing abstractive automatic text summarization in norwegian with transformers. Master's thesis, Norwegian University of Life Sciences, Ås.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Finn Årup Nielsen. 2016. Lcc. `https://github.com/fnielsen/lcc-sentiment`.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for Danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

4512–4525, Online. Association for Computational Linguistics.

Faton Rekathati. 2021. The kblab blog: Introducing a swedish sentence transformer.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

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.

Fangli Su, Yin Zhang, and Zachary Immel. 2020. Digital humanities research: interdisciplinary collaborations, themes and implications to library and information science. *Journal of Documentation*, 77(1):143–161.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Tatoeba Project Contributors. 2023. Tatoeba Corpus. `https://tatoeba.org/`. Used the version available at https://github.com/facebookresearch/LASER/tree/main/data/tatoeba/v1.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian review corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Appendix

A.1 Models

The Table 4 reference to each of the model's names denoted in the main paper, which have been shortened for clarity.

A.2 Domains Generalization

We see the performance across domains in Table 5. These results are generally in accordance with the results across tasks; showing that the e5 models along with the commercial APIs constitute the most competitive models.

A.3 Dataset Embeddings

To examine the spread and similarity of our datasets, we explore their similarity in the embedding space in Figure 4. To do so, we use one of the best-performing embedding models, embed-multilingual-v3.0. We see that certain datasets occupy distinct clusters, indicating that evaluations without these datasets would likely bias the model evaluation. Notably, we additionally see that the existing (translated) datasets within MTEB (Massive Intent and Massive Scenario) cover only a small subsection of the embedding space.

A.4 Evaluation and Metrics

This section briefly presents the tasks, their evaluation, and their metric. However, we utilize a similar implementation as MTEB to keep results comparable. Thus we refer to the original work for a more detailed introduction. We do, however, make one notable difference, that is, we allow the models to embed the tasks differently depending on the task, this is especially relevant for the e5 models, embed-multilingual-v3.0 and prompt-based models such as e5-mistral-7b-instruct.

Classification: Using the embedding model a train and a test set are embedded. Using the embedding training set a logistic classifier is trained using a maximum of 100 iterations. The model is then tested on the test set and accuracy is reported as the main metric.

Bitext Mining: The dataset consists of matching pairs of sentences, and the goal is to find the match. All matching pairs of sentences are embedded using the embedding model. Afterward, the closest match is found using cosine similarity. F1 is reported as the main metric.

Name	Reference
Self-Supervised Models	
dfm-encoder-large	`danish-foundation-models/encoder-large-v1`
+ SimCSE	`Anonymized`
xlm-roberta-large	`FacebookAI/xlm-roberta-large`
nb-bert-large	`NbAiLab/nb-bert-large`
nb-bert-base	`NbAiLab/nb-bert-base`
bert-base-swedish	`KBLab/bert-base-swedish-cased`
fasttext-cc-da	`https://fasttext.cc/docs/en/crawl-vectors.html`
fasttext-cc-nn	`https://fasttext.cc/docs/en/crawl-vectors.html`
fasttext-cc-nb	`https://fasttext.cc/docs/en/crawl-vectors.html`
fasttext-cc-sv	`https://fasttext.cc/docs/en/crawl-vectors.html`
Supervised Models	
multilingual-MiniLM-L12	`sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`
multilingual-mpnet-base	`sentence-transformers/paraphrase-multilingual-mpnet-base-v2`
labSE	`sentence-transformers/LaBSE`
sentence-bert-swedish	`KBLab/sentence-bert-swedish-cased`
e5-mistral-7b-instruct	`intfloat/e5-mistral-7b-instruct`
multilingual-e5-large	`intfloat/multilingual-e5-large`
multilingual-e5-base	`intfloat/multilingual-e5-base`
multilingual-e5-small	`intfloat/multilingual-e5-small`
translate-e5-large	Custom Implementation
sonar-dan	`facebook/SONAR`
sonar-nob	`facebook/SONAR`
sonar-nno	`facebook/SONAR`
sonar-swe	`facebook/SONAR`
Embedding APIs	
text-embedding-3-large	`https://openai.com/blog/new-embedding-models-and-api-updates`
text-embedding-3-small	`https://openai.com/blog/new-embedding-models-and-api-updates`
embed-multilingual-v3.0	`https://txt.cohere.com/introducing-embed-v3/`

Table 4: This table provides an overview, along with reference to their implementation. If a link isn't provided it denotes the name on Huggingface.

	Avg.	Fiction	Legal	News	N.-fiction	Review	Social	Spoken	Web	Wiki
Num. Datasets (→)	24	4	2	6	13	2	6	4	3	6
Self-Supervised Models										
dfm-encoder-large	41.4	44.5	69.7	31.4	33.6	56.8	42.3	57.0	29.4	31.0
+ SimCSE	**46.6**	**46.4**	72.0	**40.5**	**42.7**	58.7	**41.2**	60.7	**39.3**	37.3
xlm-roberta-large	35.3	41.5	41.3	24.9	25.3	55.9	36.2	54.4	24.4	26.5
nb-bert-large	46.0	44.0	**73.2**	38.7	42.6	**61.6**	36.1	**61.7**	30.5	**39.9**
nb-bert-base	42.1	42.6	71.8	28.7	35.1	57.6	38.4	58.7	29.0	35.0
bert-base-swedish	35.2	38.6	56.4	24.9	29.9	56.9	29.8	49.7	27.3	25.0
fasttext-cc-da	37.3	39.5	64.3	28.4	34.0	49.9	33.2	45.6	26.0	33.9
fasttext-cc-nn	35.8	38.1	64.2	24.8	33.6	47.5	29.2	43.2	24.0	35.5
fasttext-cc-nb	37.5	39.0	63.5	26.8	34.4	49.8	32.0	46.1	25.4	36.5
fasttext-cc-sv	36.0	38.3	62.7	28.0	33.3	50.9	30.1	45.8	26.6	29.3
Supervised Models										
multilingual-MiniLM-L12	50.0	43.5	68.4	43.9	49.1	59.9	45.4	57.6	43.6	41.2
multilingual-mpnet-base	53.2	44.2	72.8	47.3	52.4	64.7	49.0	59.7	45.6	43.3
labSE	50.5	49.0	71.3	41.9	48.5	61.9	48.5	57.7	48.6	44.6
sentence-bert-swedish	46.6	40.4	59.9	44.1	47.1	57.5	36.8	53.9	44.9	36.1
e5-mistral-7b-instruct	60.4	**53.7**	**77.6**	52.3	58.0	70.1	**58.0**	**64.5**	**62.1**	**57.0**
multilingual-e5-large	**60.7**	48.1	76.1	**54.5**	**58.9**	**73.5**	54.9	62.0	54.9	55.7
multilingual-e5-base	57.9	48.5	74.9	50.4	56.2	69.6	52.6	59.7	54.3	54.8
multilingual-e5-small	56.4	49.0	72.3	50.8	55.4	65.9	51.1	57.8	54.8	53.4
translate-e5-large	47.7	43.2	69.4	36.8	43.7	68.1	46.5	55.5	40.1	37.8
sonar-dan	43.4	50.2	73.5	31.0	35.7	59.1	49.2	55.5	43.0	33.1
sonar-nob	41.5	45.2	70.1	28.0	34.1	57.9	43.8	55.6	35.8	31.0
sonar-nno	41.5	46.5	71.3	28.4	33.9	58.5	44.8	56.0	37.7	30.0
sonar-swe	42.8	50.5	73.2	30.9	35.9	58.2	47.0	55.0	44.1	33.5
Embedding APIs										
text-embedding-3-large	**65.0**	**50.5**	76.1	56.1	**64.1**	72.7	**59.0**	**64.4**	**61.0**	**65.5**
text-embedding-3-small	61.0	50.2	75.9	54.0	61.2	66.6	55.3	61.2	58.1	60.7
embed-multilingual-v3.0	64.1	49.2	**76.6**	**56.2**	63.5	**75.2**	57.1	63.3	57.9	63.6

Table 5: Performance across domains in SEB. The best score in each model category is highlighted in bold. We only include domains that contain at least two datasets. Additional model evaluation can be found on the public Dashboard[8].

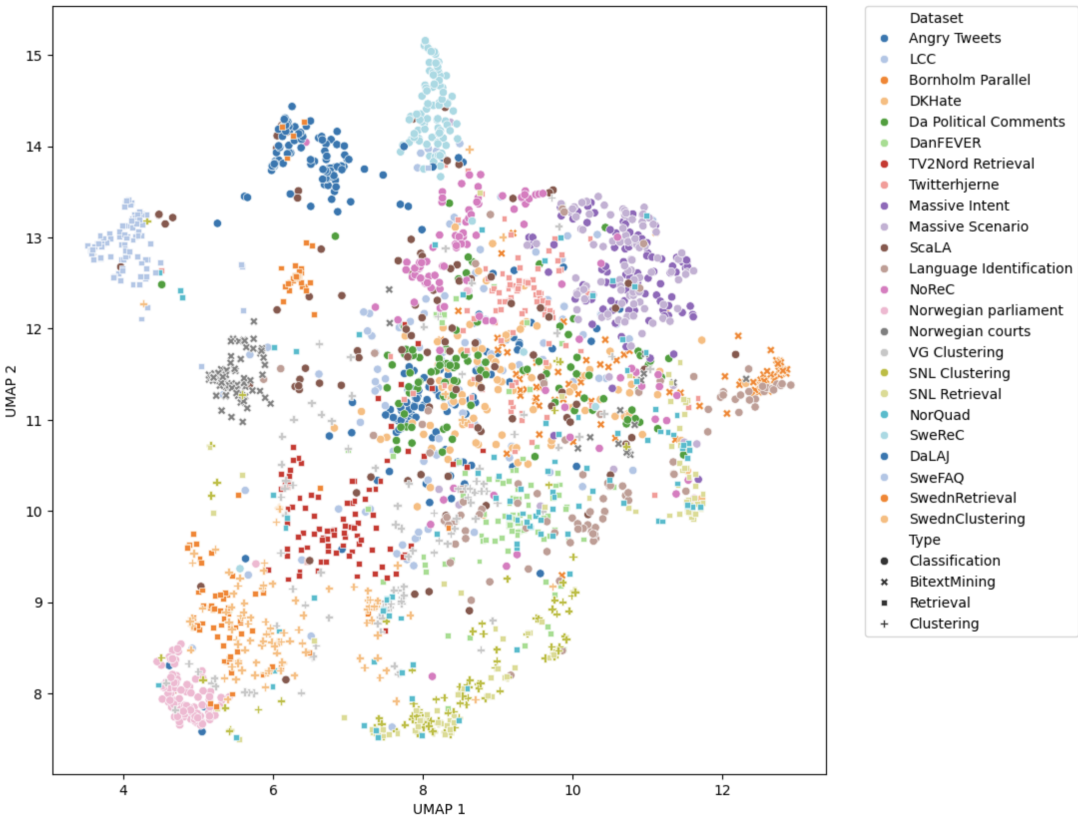

Figure 4: The embeddings of 100 randomly sampled documents from each task, embedded using embed-multilingual-v3.0 and projected using a UMAP projection. The project uses the cosine metrics but otherwise default parameter values.

Clustering The dataset consists of documents attached with a label, e.g., a denoted category such as "sports." The goal is the correctly cluster the documents into similar clusters as the labels. All documents are embedded, and a mini-batch k-means model (batch size 32 and k equal to the number of unique labels in the dataset) is trained on the embeddings. The V measure is used as is reported as the main metric, as the permutation of labels does not affect the score.

Retrieval: The dataset consists of a corpus, queries as well as a mapping between the queries and their relevant documents. The goal is to retrieve these relevant documents. Both queries and documents are embedded using the model. We allow these to be embedded differently depending on the model. For each query, the corpus documents are ranked using a similarity score, and nDCG@10 is reported as the main metric.

A.5 Datasets Construction

This section briefly describes the construction of the tasks.

Classification: As all the classification datasets are derived from existing datasets, no additional processing is done to these except to limit the size of excessively large datasets.

Bitext Mining: Similar to the classification, these datasets already existed as paired datasets. With the Norwegian Courts being extracted from OPUS (Tiedemann, 2012) and Bornholm Parallel being derived from (Derczynski and Kjeldsen, 2019).

Clustering: For clustering, we construct the dataset based on existing datasets of news or encyclopedic corpora (Navjord and Korsvik, 2023; Berdicevskis et al., 2023) using their attached categories. The SNL and VG datasets (Navjord and Korsvik, 2023) contain a hierarchy of labels; here, we subjectively chose a meaning level and validated that it led to a meaningful separation in performance – using either too many or too few levels would to either 1-2 clusters or clusters consisting of only 2-3 documents.

Similar to the classification, these datasets already existed as paired datasets. With the Norwegian Courts being extracted from OPUS (Tiedemann, 2012) and Bornholm Parallel being derived from (Derczynski and Kjeldsen, 2019).

Retrieval: For the construction of the retrieval datasets, we used either question and answer datasets (e.g., NorQuad (Ivanova et al., 2023)) or news summarization datasets (e.g., (Berdicevskis et al., 2023)). For the question and answer we specified the questions as queries and the answers as the corpus. A correct answer was considered to be a properly retrieved document. For the summaries, we considered the headline as the query and both the summaries and the articles as the corpus. Matching summaries and articles were considered properly retrieved documents.

A.6 Datasets Statistics

Table 6 contains an overview of each of the datasets in SEB, including a short description, descriptive statics, task formalization, and domains as defined by (Nivre et al., 2017).

Dataset	Description	Main Score	Langs	Type	Domains	N. Docs	Avg. Length
Angry Tweets (Pauli et al., 2021)	A sentiment dataset with 3 classes (positiv, negativ, neutral) for Danish tweets	Accuracy	da	Classification	social	1047	156.15 (82.02)
Bornholm Parallel (Derczynski and Kjeldsen, 2019)	Danish Bornholmsk Parallel Corpus. Bornholmsk is a Danish dialect spoken on the island of Bornholm, Denmark.	F1	da, da-bornholm	BitextMining	poetry, wiki, fiction, web, social	1000	44.36 (41.22)
DKHate (Sigurbergsson and Derczynski, 2020)	Danish Tweets annotated for Hate Speech either being Offensive or not	Accuracy	da	Classification	social	329	88.18 (68.30)
Da Political Comments	A dataset of Danish political comments rated for sentiment	Accuracy	da	Classification	social	7206	69.60 (62.85)
DaLAJ (Berdicevskis et al., 2023)	A Swedish dataset for linguistic acceptability. Available as a part of Superlim	Accuracy	sv	Classification	fiction, non-fiction	888	120.77 (67.95)
DanFEVER (Nørregaard and Derczynski, 2021)	A Danish dataset intended for misinformation research. It follows the same format as the English FEVER dataset.	NDCG@10	da	Retrieval	wiki, non-fiction	8897	124.84 (168.53)
LCC (Nielsen, 2016)	The Leipzig corpora collection, annotated for sentiment	Accuracy	da	Classification	legal, web, news, social, fiction, non-fiction, academic, government	150	118.73 (57.82)
Language Identification (Haas and Derczynski, 2021)	A dataset for Nordic language identification.	Accuracy	da, sv, nb, nn, is, fo	Classification	wiki	3000	78.23 (48.54)
Massive Intent (FitzGerald et al., 2022)	The intent task within MASSIVE corpus translated for Scandinavian languages	Accuracy	da, nb, sv	Classification	spoken	15021	34.65 (16.99)
Massive Scenario (FitzGerald et al., 2022)	The scenario task within MASSIVE corpus translated for Scandinavian languages	Accuracy	da, nb, sv	Classification	spoken	15021	34.65 (16.99)

Dataset	Description	Main Score	Langs	Type	Domains	N. Docs	Avg. Length
NoReC (Velldal et al., 2018)	A Norwegian dataset for sentiment classification on review	Accuracy	nb	Classification	reviews	2048	89.62 (61.21)
NorQuad (Ivanova et al., 2023)	Human-created question for Norwegian Wikipedia passages.	NDCG@10	nb	Retrieval	non-fiction, wiki	2602	502.19 (875.23)
Norwegian courts (Tiedemann, 2012)	Nynorsk and Bokmål parallel corpus from Norwegian courts.	F1	nb, nn	BitextMining	legal, non-fiction	456	82.11 (49.48)
Norwegian parliament	Norwegian parliament speeches annotated with the party of the speaker (‘Sosialistisk Venstreparti‘ vs ‘Fremskrittspartiet‘)	Accuracy	nb	Classification	spoken	2400	1897.51 (1988.62)
SNL Clustering (Navjord and Korsvik, 2023)	Webscrabed articles from the Norwegian lexicon 'Det Store Norske Leksikon'. Uses article's categories as clusters.	V measure	nb	Clustering	non-fiction, wiki	2048	1101.30 (2168.35)
SNL Retrieval (Navjord and Korsvik, 2023)	Webscrabed articles and ingresses from the Norwegian lexicon 'Det Store Norske Leksikon'.	NDCG@10	nb	Retrieval	non-fiction, wiki	2600	1001.43 (2537.83)
ScaLA (Nielsen, 2023)	A linguistic acceptability task for Danish, Norwegian Bokmål Norwegian Nynorsk and Swedish.	Accuracy	da, nb, sv, nn	Classification	fiction, news, non-fiction, spoken, blog	8192	102.45 (55.49)
SweFAQ (Berdicevskis et al., 2023)	A Swedish QA dataset derived from FAQ	NDCG@10	sv	Retrieval	non-fiction, web	1024	195.44 (209.33)
SweReC (Nielsen, 2023)	A Swedish dataset for sentiment classification on review	Accuracy	sv	Classification	reviews	2048	318.83 (499.57)
SwednClustering (Berdicevskis et al., 2023)	News articles from the Swedish newspaper Dagens Nyheter (DN) collected during the years 2000–2020. Uses the category labels as clusters.	V measure	sv	Clustering	non-fiction, news	2048	1619.71 (2220.36)

Dataset	Description	Main Score	Langs	Type	Domains	N. Docs	Avg. Length
SwednRetrieval (Berdicevskis et al., 2023)	News articles from the Swedish newspaper Dagens Nyheter (DN) collected during the years 2000–2020.	NDCG@10	sv	Retrieval	non-fiction, news	3070	1946.35 (3071.98)
TV2Nord Retrieval	News Article and corresponding summaries extracted from the Danish newspaper TV2 Nord.	NDCG@10	da	Retrieval	news, non-fiction	4096	784.11 (982.97)
Twitterhjerne (Holm, 2024)	Danish question asked on Twitter with the Hashtag #Twitterhjerne ('Twitter brain') and their corresponding answer.	NDCG@10	da	Retrieval	social	340	138.23 (82.41)
VG Clustering (Navjord and Korsvik, 2023)	Articles and their classes (e.g. sports) from VG news articles extracted from Norsk Aviskorpus.	V measure	nb	Clustering	non-fiction, news	2048	1009.65 (1597.60)

Table 6: Tasks available in SEB. The average length is specified in characters. Values in parentheses denote the standard deviation.

A.7 Results per Task

In the following figure, we see an overview of all of the results of the benchmark for each task for the selected models. To get an up-to-date overview, check out the online dashboard.

Model	Average Score	Average Rank	Angry Tweets	Bornholm Parallel	DKHate	Da Political Comments	DaLAJ	Dan-FEVER	LCC	Language Identification	Massive Intent	Massive Scenario	NoReC	NorQuad	Norwegian courts	Norwegian parliament	SNL Clustering	SNL Retrieval	ScaLA	SweFAQ	SweReC	SwednClustering	SwednRetrieval	TV2Nord Retrieval	Twitterhjerne	VG Clustering
multilingual-e5-base	57.9	11.4	56.3	33.2	63.8	36.3	49.8	40.1	60.3	75.9	61.0	67.9	59.0	21.9	89.5	59.6	63.9	94.2	50.5	69.5	80.2	10.9	60.7	92.7	65.4	27.2
multilingual-e5-small	56.4	12.6	56.2	37.1	62.4	34.7	50.0	38.3	58.5	72.1	56.6	64.4	54.5	17.5	86.0	60.0	63.4	91.7	50.3	68.7	77.4	16.4	58.3	90.4	57.4	30.9
multilingual-mpnet-base	53.2	14.6	54.9	18.2	63.8	41.3	50.0	37.2	58.4	41.6	63.4	70.9	56.1	38.7	87.3	54.6	61.9	62.5	50.0	60.4	73.4	9.0	60.8	78.4	57.6	27.1
nb-bert-large	46.0	16.7	52.1	4.5	62.1	35.6	50.9	25.8	56.3	85.3	58.2	61.7	55.5	17.2	90.1	62.6	67.1	39.7	64.2	30.7	67.7	11.7	21.4	50.3	6.0	28.2
LaBSE	50.5	17.6	52.1	45.6	62.7	38.7	49.8	34.2	50.1	35.4	58.6	65.2	51.2	30.5	92.6	56.8	62.7	59.3	50.4	50.1	72.5	5.5	50.4	76.3	41.7	18.7
multilingual-MiniLM-L12	50.0	18.0	50.9	19.7	59.1	37.4	50.1	36.5	54.3	42.5	57.5	66.1	49.9	34.7	82.4	56.6	61.9	52.1	50.0	56.9	70.0	6.8	52.8	73.3	51.2	26.2
dfm-encoder-large (SimCSE)	46.6	19.2	54.4	15.9	63.2	38.5	50.0	36.9	58.1	76.0	59.6	64.1	50.5	10.7	86.0	57.7	63.0	21.6	61.5	43.8	67.0	3.9	24.9	80.8	17.0	13.7
translate-e5-large	47.7	19.8	54.9	17.6	59.8	34.8	50.2	34.5	55.0	43.8	55.8	63.0	55.9	13.9	83.7	53.1	61.5	55.5	50.0	47.8	80.3	5.9	33.0	62.5	56.7	14.6
nb-bert-base	42.1	20.7	52.1	9.9	61.7	34.3	50.3	21.5	51.4	84.7	57.1	61.5	51.3	10.8	92.2	57.4	60.4	22.7	58.8	25.6	63.9	9.0	18.0	9.3	21.1	26.0
sentence-bert-swedish-cased	46.6	21.0	44.5	14.1	59.4	28.5	50.1	30.1	47.2	51.4	51.6	58.4	43.5	10.1	72.6	55.7	65.8	45.3	50.1	73.3	71.4	15.5	70.6	55.8	26.9	25.5
sonar-dan	43.4	22.1	48.2	47.1	70.4	33.7	50.0	24.2	53.1	46.6	54.9	62.7	50.6	7.3	93.9	54.0	44.9	28.7	50.5	28.9	67.7	2.1	22.8	45.6	42.8	11.9
sonar-swe	42.8	23.2	47.3	48.1	70.0	31.8	50.1	24.1	53.1	45.8	54.2	61.1	49.9	7.0	93.3	54.4	47.0	28.8	50.5	31.2	66.4	3.3	23.2	47.2	31.6	7.8
dfm-encoder-large	41.4	23.2	53.8	11.6	60.1	37.1	50.4	24.1	57.3	77.7	54.3	56.3	48.3	3.0	82.0	58.8	62.7	6.7	58.6	19.1	65.2	4.6	6.8	47.7	33.7	13.4
sonar-nob	41.5	24.0	47.9	33.1	69.7	32.5	50.1	22.2	46.9	49.2	54.4	61.9	48.7	6.5	93.3	55.4	44.4	30.8	50.8	27.5	67.0	2.3	17.9	41.3	32.7	8.9
sonar-nno	41.5	24.2	48.1	36.6	68.8	32.4	50.1	22.0	48.4	44.7	56.3	62.5	48.5	5.5	94.3	54.7	42.9	28.1	50.8	28.1	68.6	1.1	21.2	41.0	34.3	7.8
xlm-roberta-large	35.3	24.5	51.7	4.3	60.2	31.9	52.5	10.6	48.7	81.3	48.8	50.8	44.6	2.0	33.9	57.7	59.2	1.7	60.3	20.0	67.2	10.7	9.2	6.1	20.4	14.4
bert-base-swedish-cased	35.2	27.6	44.6	6.6	55.5	28.5	51.8	16.0	41.2	62.4	42.2	44.1	43.9	1.0	71.5	57.6	60.0	4.2	54.9	34.0	69.8	8.1	25.0	9.7	2.6	10.6
fasttext-cc-nb-300	37.5	28.1	46.0	7.6	52.7	29.0	50.1	24.8	48.3	74.2	34.2	43.0	40.9	7.7	78.8	57.3	59.8	44.7	50.0	20.4	58.8	2.0	17.3	32.3	8.4	10.8
fasttext-cc-sv-300	36.0	29.4	42.7	7.1	55.8	27.3	50.2	23.1	45.9	60.3	34.3	42.7	37.8	5.5	79.6	56.1	53.6	26.4	50.1	26.8	64.1	4.8	31.8	27.6	1.8	7.7
fasttext-cc-da-300	37.3	29.6	47.3	7.1	53.6	29.9	50.0	27.0	50.9	71.6	34.3	42.3	39.8	6.6	77.7	55.5	56.4	34.7	50.1	19.9	60.0	2.6	17.1	43.0	10.4	6.5
fasttext-cc-nn-300	35.8	30.2	42.4	9.5	51.9	27.7	50.1	23.4	42.6	71.6	29.5	35.9	37.6	6.9	85.8	57.2	56.3	45.2	50.1	19.9	57.5	3.3	16.3	29.8	1.1	6.6
text-embedding-3-large	65.0	6.4	57.8	43.3	70.2	43.4	50.0	39.6	58.1	79.7	69.6	76.2	61.6	68.1	94.2	61.4	65.2	97.1	50.4	81.6	83.7	16.1	82.2	95.2	81.1	34.9
embed-multilingual-v3.0	64.1	7.3	58.7	35.6	68.8	43.4	50.0	41.0	60.4	78.7	67.8	74.7	66.1	60.9	92.9	60.0	69.8	95.8	50.7	77.7	84.4	15.0	80.0	95.4	75.8	35.8
multilingual-e5-large	60.7	8.9	57.7	29.6	66.2	39.7	49.9	40.5	61.7	80.2	64.9	71.4	63.5	25.6	90.5	60.3	62.8	95.5	51.2	73.3	83.4	12.0	79.2	95.4	74.4	27.9
text-embedding-3-small	61.0	9.4	55.6	41.0	65.6	39.8	50.1	39.1	59.4	67.9	63.9	71.9	55.7	57.6	92.4	58.8	66.0	92.7	50.3	73.9	77.4	14.4	73.5	92.0	70.3	34.5
e5-mistral-7b-instruct	60.4	9.4	58.4	50.5	64.5	39.7	50.3	38.2	63.9	65.2	71.0	76.0	60.2	27.5	91.2	60.7	66.3	94.3	50.2	72.0	79.9	11.2	67.6	91.2	71.1	29.5