Comparing Tools for Sentiment Analysis of Danish Literature from Hymns to Fairy Tales: Low-Resource Language and Domain Challenges

Pascale Feldkamp
Center for Humanities Computing
Aarhus University
pascale.moreira@cc.au.dk

Jan Kostkan
Center for Humanities Computing
Aarhus University
jan.kostkan@cas.au.dk

Ea Lindhardt Overgaard
School of Communication and Culture
Aarhus University
elt@cc.au.dk

Mia Jacobsen
Center for Humanities Computing
Aarhus University
miaj@cas.au.dk

Yuri Bizzoni
Center for Humanities Computing
Aarhus University
yuri.bizzoni@cc.au.dk

Abstract

While Sentiment Analysis has become increasingly central in computational approaches to literary texts, the literary domain still poses important challenges for the detection of textual sentiment due to its highly complex use of language and devices – from subtle humor to poetic imagery. Furthermore, these challenges are only further amplified in low-resource language and domain settings. In this paper we investigate the application and efficacy of different Sentiment Analysis tools on Danish literary texts, using historical fairy tales and religious hymns as our datasets. The scarcity of linguistic resources for Danish and the historical context of the data further compounds the challenges for the tools. We compare human annotations to the continuous valence scores of both transformer- and dictionary-based Sentiment Analysis methods to assess their performance, seeking to understand how distinct methods handle the language of Danish prose and poetry.

1 Introduction and related works

Sentiment Analysis (SA) is a highly popular field in Computational Linguistics and NLP, as it attempts to interpret the sentimental and emotional aspects of texts, with applications that range from consumer review analysis (Tsao et al., 2018) to social media monitoring (Bollen et al., 2011; Asur and Huberman, 2010). It is, moreover, an increasingly central method for computational literary studies research as well (Rebora, 2023), where it has found popular applications to explore the narrative development (Zehe et al., 2016) or visualizing "sentiment arcs" of novels (i.e., the sequential highs and lows of valence throughout a narrative) (Jockers, 2014; Reagan et al., 2016). The sentiment arcs of novels – after applying a detrending technique to abstract from the noisy signal of raw valence scores – have also been used to assess, for example, the connection between narrative dynamics and reader appreciation (Bizzoni et al., 2023).

Still, the relation between sentiment arcs extracted with SA tools and actual reader experience remains understudied – both in their raw and detrended forms. Though recent studies of narrative sentiment arcs, like that of Elkins (2022), go some way in comparing various approaches to SA, they either do not contrast SA tools against a human gold standard at a granular level or have focused on single case studies (Bizzoni and Feldkamp, 2023).

Partially this is due to the very complexity of the literary domain. Literary texts often aim to evoke rather than explicitly communicate; operate at multiple narrative levels (Jakobson, 2010 (1981; Rosenblatt, 1982; Booth, 1983); make high use of ambiguity and poetic devices; offer several interpretations; and have been shown to rely on specific linguistic registers to evoke affective reactions (Bizzoni and Feldkamp, 2024).[1] For these reasons, SA tools might be more effective in other domains (Alantari et al., 2022; Elsahar and Gallé, 2019; Ohana et al., 2012; Bowers and Dombrowski, 2021) than the literary, although some studies have suggested that Transformer-based models might be able to bridge the gap and perform better on literary

[1] Naturally, these phenomena extend outside the literary domain as well (Rentoumi et al., 2009), for example, tweets using irony or figurative language likely effect diverging interpretations (Sandri et al., 2023; Stengel-Eskin et al., 2021).

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 186–199

or poetic material as well (Schmidt et al., 2021).

Beyond domain-specificity, an obvious obstacle to a wider use of SA for literature is the issue of multilinguality. The majority of research in SA – both in more general NLP and in the literary domain – has concentrated on well-resourced languages like English (Ribeiro et al., 2016). Once again, Transformer-based architectures able to generalize across multiple languages (Devlin et al., 2019) have helped reduce the gap, and multilingual transformers hold a significant promise for cross-lingual SA (Elkins, 2022), but language- and culture-related biases from English pretraining have been shown to impact the performance of transfer learning on low-resource languages (De Bruyne et al., 2022; Papadimitriou et al., 2023; Xu et al., 2022).

When it comes to Danish specifically, the main dictionary-based SA tools – Afinn, Sentida, and Danish Sentiment Lexicon (DSL) – have been shown to perform comparably across domains (Schneidermann and Pedersen, 2022), with Sentida in particular, showing a strong correlation with human judgments for both fiction and social media (Lauridsen et al., 2019). While such dictionaries appear to show a consistent performance for Danish SA, they are not widely tested at a fine-grained level, nor on historical Danish. Assessing the performance of models on historical Danish and Norwegian literary texts, Allaith et al. (2023) found that multilingual transformer models outperformed both fine-tuned models and classifiers based on lexical resources in the target language, which aligns with the findings of Schmidt et al. (2021) and Schmidt and Burghardt (2018) for historical German drama.

With the present study, we seek to examine two main issues: i) the challenge for SA models of understanding sentiment in *historical* literary texts – both prose (fairy tales) and poetry (religious hymns); ii) the challenge of applying SA models on fiction written in under-resourced languages like Danish. We evaluate how different SA tools – transformer-based and dictionary-based approaches – perform on the literary texts compared to a human gold standard.[2] In addition, we apply three English-based methods widely used for literary SA (Bowers and Dombrowski, 2021; Elkins and Chun, 2019; Bizzoni et al., 2023) on text that was Google-translated, as a point of comparison for the performance of Danish-based tools. Finally, we examine SHAP-scores of the best-performing transformer-based method to gauge differences between transformer- and dictionary-based methods.

2 Methodology

2.1 Datasets

We use two different datasets: (i) three literary fairy tales by Hans Christian Andersen and (ii) a collection of Danish religious hymns. We selected these datasets to provide a historical while rich and varied set of Danish literature, taking both narrative and poetic complexity as well as their cultural significance into consideration.[3]

The HCA dataset is larger than the hymns dataset by number of words (Table 1) – but not by number of annotations (fairy tales were annotated on a sentence- and hymns on a verse-basis). Both datasets are from within the period 1798–1873, which is additionally challenging for models predominantly based/pretrained on modern Danish.

	Texts	V/S	Words	$\bar{x}$ V/S	Period
HCA	3	791	18,910	263.7	1837-1847
Hymns	65	1,914	10,303	32.9	1798-1873

Table 1: The **HCA** and the **Hymns** datasets: The total number of verses or sentences (V/S) and words per dataset, and mean ($\bar{x}$) number of verses or sentences per text.

Literary Fairy Tales The HCA dataset includes three of Andersen's most known fairy tales: "The Little Mermaid" (1837), "The Ugly Duckling" (1844), and "The Shadow" (1847)(CCLM, 2003).[4] These texts are emblematic of Danish cultural heritage and literary tradition, known for a simple but involving narrative and memorable character representations. Andersen's fairy tales often contain multiple layers of meaning and sentiment, ranging from joy and wonder to sadness and introspection, while keeping an essential simplicity, both stylistically and in the narrative arc (Lundskær-Nielsen, 2014; Alm and Sproat, 2005), which makes them

[2]The annotated resource is available for further studies at: https://github.com/centre-for-humanities-computing/Danish_literary_sentiment/

[3]Andersen's production being arguably the most central in Danish literary heritage (Ringgaard and Thomsen, 2017), while hymns of N.F.S. Grundtvig (also included here) are less internationally known but equally significant in shaping the national cultural identity (Nielsen, 2020).

[4]Spelling has been modernized in these texts editions, though vocabulary has not been significantly changed.

an ideal case for testing sentiment analysis tools on literary Danish.

Religious Hymns To further create a literary challenge for tools, we used a hymns dataset, comprising 65 Danish religious hymns around the 19th century,[5] where each verse is coupled with its modernized Danish version.[6] The hymns are characterized by a more structured formality and an archaic and poetic language, especially in the original versions – for example, the use of the latinized "est" for "is" ("*er*"). The inclusion of both original and modernized texts allows us to observe whether language evolution might significantly affect Danish SA. Hymns are challenging for SA tools due to the poetic and figurative language, subtle emotional tones, as well as their cultural and religious contexts – especially Christian values and symbolic structures of meaning (Skovsted et al., 2019; Nielsen, 2020). Finally, while the prosaic fairy tales are divided into sentences, verses were chosen as the unit of analysis for the hymns, seeing that the verse constitutes the building block of poetry more than the sentence. A syntactically sound sentence might thus not be present in every verse, so verses may be syntactically simpler but semantically more challenging, which may further confound sentiment annotation (both human and automatic).

We selected Andersen's fairy tales and Danish hymns to challenge and evaluate sentiment analysis tools across two very different, but highly representative, types of literary texts. Andersen's tales have narratives and emotional depth, but use standard prose linguistic structures - so they will test the models' ability to handle complex emotional narratives. In contrast, the hymns rely on poetic expression and do not represent a story but rather a non-narrative message. They provide a test case for the models' sensitivity to subtler, less structured, sentiment evocation.

[5]From 1798 (n=35), 1857 (n=17) and 1873 (n=13). Note that the years refer to the publication date of three official church hymn collections.

[6]Two literary scholars modernized the original Danish prompting ChatGPT 3.5 (prompt: "Oversæt til moderne dansk retstavning", i.e. "translate to modern Danish spelling".), and subsequently validated each output verse against the original. The date for this was May 20, 2024.

2.2 English Translations

We obtained translations of sentences and verses via google-translate (not manually validated).[7] We used these translations in combination with two English dictionary-based systems and the RoBERTa base xlm multilingual (which we also apply to the original Danish) as a raw baseline for comparison to systems developed in and for Danish.

2.3 Human annotation

Human annotators (n=3) read H.C. Andersen's fairytales from beginning to end and scored *each sentence* on a 0 to 10 valence scale:[8] 0 signifying the lowest, and 10 the highest valence.[9] For the hymns, annotators (n=2) read and scored *each verse line* on the same scale. The valence score was intended to represent the sentiment expressed by the sentence and verse. The annotators were instructed to avoid rating how a sentence or verse made them feel and to try to report only on the sentiments actually embedded in the sentence, i.e., to think about the valence of the individual sentence and verse, without overthinking the story's/hymn's narrative to reduce contextual interpretation.

It is worth noting that humans rarely reach an agreement higher than 80% (or 0.80 Krippendorff's α) on non-fiction texts for tasks like positive/neutral/negative discrete tagging (Wilson et al., 2005) or continuous scale polarity annotation (Batanović et al., 2020). In our case, detrending the annotators' scores (see Section 2.4.4) always improved the Inter Annotator Reliability (Table 2), which might be seen as a natural effect of smoothing time-series (removing outliers). An example of detrended arcs of the annotators' individual and mean scores – the latter of which is used to compare systems' scores – is visualized in Fig. 1.

2.4 Automatic annotation

We used several SA models for Danish, transformer- and dictionary-based, to score the texts for valence.

[7]We used the deep translator package in python to retrieve google-translated sentences: https://pypi.org/project/deep-translator/ Translations were retrieved on May 20, 2024.

[8]Sentences were tokenized using the nltk tokenize package: https://www.nltk.org/api/nltk.tokenize.html

[9]Annotators were all native Danish speakers, two with a background in literary studies (MA, PhD) and one from cognitive science (MA). The two annotators of the hymns (MA and PhD of literature) had domain knowledge in 19th century Scandinavian literature and historical religious hymns.

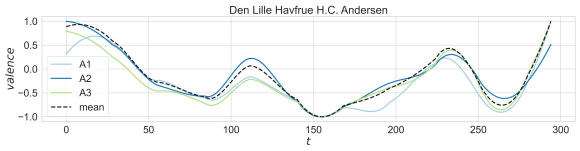


Figure 1: Sentiment arcs of **The Little Mermaid** after detrending annotators' values. The black line represents the mean annotator score.

2.4.1 Dictionary- and rule-based methods

Dictionary-based methods (that are usually rule-based as well) – meaning tools that use a predefined dictionary to assign basic valence to words – remain popular especially in humanities research, due to their transparency and versatility. Moreover, they seem to perform well (Bizzoni and Feldkamp, 2023) – even on so-called "nonlinear" narratives (Richardson, 2000; Elkins and Chun, 2019) although they appear to do poorly on a word-basis (Reagan et al., 2017). Our chosen models were:

Afinn: valence dictionary without rules, extracted from twitter-data and various open sources.[10] The dictionary contains many inflections of the same lemma. Valence scores range from -5 to +5.

Sentida: a rule-based system inspired by the English VADER (observes negations, adverb modifiers, etc.).[11] Sentida combines the Afinn dictionary with the 10,000 most frequent Danish lemmas, that were manually annotated by the authors (Lauridsen et al., 2019). Upon inference, it relies on stemming to find matching dictionary items. Valence scores range from -5 to +5.

Asent: a rule-based system, using the Afinn dictionary by default, while adding rules (e.g. negations, modifiers, intensifiers, etc.).[12] Valence scores range from -1 to +1.

Score normalization For comparing models on raw scores, we maintained the different eays of scaling in each dictionary-based method. For detrending the time series, however, we normalized all scales – including the human annotation scale – to the range -1 to +1.[13]

2.4.2 Transformer-based methods

More recent Transformer-based approaches have found application both in Danish and as multilingual models, and have shown both potential and pitfalls in SA for literary texts (Elkins, 2022). We chose to use all off-the-shelf models currently developed for Danish SA and a widely used multilingual model, RoBERTa xlm (Conneau et al., 2020):[14]

Senda: was developed specifically for Danish.[15] It was built on the Roberta architecture, pretrained on a large corpus of Danish texts.

Alexandra institute sentiment base:[16] is another example of a Danish-oriented transformer that has been fined-tuned for SA tasks. It is hosted by the Alexandra Institute.

RoBERTa base xlm multilingual:[17] was trained using the cross-lingual language training approach, that is supposed to enhance its ability of understanding and processing tens of different languages by transferring its learned skills – in other words, by using what it has learned from one language to help it in another language. Its ability to transfer learning across languages might potentially allow it to generalize more powerfully on sentiment analysis, but it could also hinder its ability to deal with language-specific expressions, especially in unusual domains.

Score transformation Note that we converted the categorical Transformer output to continuous SA scores by using the confidence score of labels as a proxy for sentiment intensity. If the model classifies a sentence as *positive* with a confidence of, for example, 0.89, we interpret it as a valence score of +0.89 for this sentence, and so on. Note that we converted scores of the *neutral* category to neutral (0), also seeing that most human scores fall into the vicinity of neutral (5 on the human 0-10 annotation scale).[18][19]

2.4.3 English-based models

To compare Danish tools to English tools as a baseline, we used the Google translated sentences (see section 2.2), applying often used English-language

[10] https://github.com/fnielsen/afinn
[11] https://github.com/Guscode/Sentida
[12] https://github.com/KennethEnevoldsen/asent
[13] We used the MinMaxScaler-approach for normalization.
[14] We maintained all presets as the default when applying these models, so that the hyperparameters are as specified in the documentation of the individual model (see the model hyperlinks).
[15] https://huggingface.co/larskjeldgaard/senda
[16] https://huggingface.co/alexandrainst/da-sentiment-base
[17] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
[18] For the distribution of scores, see Appendix (Fig. 6a).
[19] Bizzoni and Feldkamp (2023) similarly used this method for converting discrete transformer output to continuous scores.

tools. We chose the two dictionary- and rule-based models **VADER** and **Syuzhet**, because of their popularity and use in literary SA studies (Allaith et al., 2023; Bizzoni et al., 2022b; Bizzoni and Feldkamp, 2023), and the same **RoBERTA** multilingual model as applied on the Danish texts (see above), due to observed good performance on literary prose in Bizzoni and Feldkamp (2023). All of these were applied to Google-translated sentences which had not been manually checked for accuracy.

2.4.4 Arcs and Detrending

In the analysis of sentiment within literary texts, the consideration of narrative arcs has been central (Rebora, 2023), particularly for texts with a clear story progression like fairy tales, where studies have used detrending methods to gauge the role of sentiment dynamics for reader appreciation (Bizzoni et al., 2022a). However, for other types of literature such as hymns, which do not exhibit explicit story development, narrative arcs are not as apt as an analytical framework. For this reason, we consider both the raw and detrended sentiment arcs of the fairy tales in our dataset, but *do not detrend the hymns*. For the fairy tales, we examine whether detrending can improve the correlation between scores and human annotations. The detrended scores are derived through a polynomial fit of the original data, designed to smooth out the noise and highlight the overall narrative shape.[20] The detrending process allows observation of the underlying emotional trajectory of the story without the interference of short-term fluctuations, providing a clearer view of how sentiments evolve.

3 Results

3.1 Human annotation

We report a relatively high inter-rater reliability (IRR), with a correlation (Spearman's *rho*) between their scores of 0.726 for the hymns.[21] For the fairy tales we find an average correlation of 0.64 – non-detrended; detrended annotator scores have a correlation coefficient > 0.80 (Table 2). As mentioned, higher agreement for detrended values is an effect of smoothing values (removing outliers) and suggests that annotators agree on the overall shape of the narrative when abstracting from the granular level. IRR is high, especially in the case of hymns (considering the fragmentariness of the verses) and considering that humans often have low agreement for sentiment annotation, not least continuous-scale annotation.[22]

	Spearman's *rho* ($\bar{x}$)	**Krippendorff's α**
Mermaid	0.80 (0.94)	0.85 (0.91)
Duckling	0.47 (0.89)	0.65 (0.90)
Shadow	0.65 (0.80)	0.76 (0.78)
Hymns	0.73 (-)	0.72 (-)

Table 2: Inter Rater Reliability between annotators (n=3) in the fairy tales, using the mean Spearman correlation coefficient ($p<0.01$) – with Krippendorff's Alpha for reference. Correlation between the annotators' non-detrended values and detrended values (in parenthesis).

3.2 Sentiment Analysis on Andersen's Fairy Tales

The sentiment scoring of H.C. Andersen's fairy tales *The Ugly Duckling*, *The Little Mermaid* and *The Shadow* appears quite challenging for both dictionary- and transformer-based models. Considering raw (non-detrended) scores, the transformer-based models generally perform better than dictionary-based tools across all three stories (Table 3). Notably, *The Ugly Duckling* shows the highest Spearman correlation with RoBERTa (0.58) and *The Little Mermaid* with Asent (0.54) and Sentida (0.51). Human annotations of *The Shadow* also appear more aligned to RoBERTa again, achieving a correlation of 0.56. Still, it should be noted that RoBERTa does not perform consistently (i.e., in the case of *The Little Mermaid*) where dictionary-based Syuzhet on Danish-English Google translations are performing comparably and more consistently. Notably, the best and most consistently performing system appears to be the RoBERTa applied to Google translations.

When considering detrended scores improvement is evident across most models. Note that the correlation (and Krippendorrf's α) also improves when human scores are detrended (Table 2). *The*

[20] We use Nonlinear Adaptive Filtering technique to detrend arcs. For more on this method, see Jianbo Gao et al. (2010) as well as the implementation on narrative fiction in Hu et al. (2021).

[21] We report the Spearman correlation coefficient here. As annotators operated within a continuous valence spectrum, divided into ten categories, we find that correlation measures more clearly reflect the values' direction and nuance (parallelity vs exactness), compared to categorical inter-annotator agreement measures. We provide Krippendorff's α for reference, where the level of measurement was considered interval.

[22] For a continuous sentiment annotation task similar to the one presented here – albeit on modern fiction – Bizzoni and Feldkamp (2023) report a Spearman correlation between annotators (n=2) of 0.624.

	Afinn	Sentida	Asent	Alex.in.	Senda	RoB	VADER	Syuzhet	RoB
Duckling	0.29	0.44	0.28	0.50	0.45	**0.58**	0.42	0.50	**0.57**
Mermaid	0.37	0.51	**0.54**	0.49	0.37	0.38	0.49	0.51	**0.52**
Shadow	0.38	0.34	0.39	0.43	0.28	**0.56**	0.51	0.47	**0.63**
Average	0.35	0.43	0.40	0.47	0.37	**0.51**	0.47	0.49	**0.57**
Duckling (D.)	0.41	0.18	0.42	0.65	0.55	**0.67**	0.32	0.46	**0.62**
Mermaid (D.)	0.70	0.73	0.72	0.71	0.01	**0.75**	**0.81**	0.63	0.71
Shadow (D.)	0.39	0.53	0.39	0.40	0.25	**0.70**	0.42	0.45	**0.82**
Average (D.)	0.50	0.48	0.51	0.59	0.27	**0.71**	0.52	0.51	**0.72**

Table 3: Spearman correlation between **raw** (above) / **detrended arcs** (below), i.e., between raw/detrended system scores and raw/detrended human mean scores. Dictionary and rule-based systems (left), transfomer-based systems (middle) and three English systems' scores on Google-translated sentences included as a baseline (right). Note that RoBERTa (RoB) on the right was used on translated sentences, and RoBERTa on the left on the original Danish sentences. Best performing Danish tools in bold, best baseline in green. Note that although correlations on detrended arcs seem high, on *The Little Mermaid*, all correlations (Spearman's *rho*) between annotators' detrended arcs have a Spearman correlation >0.93.

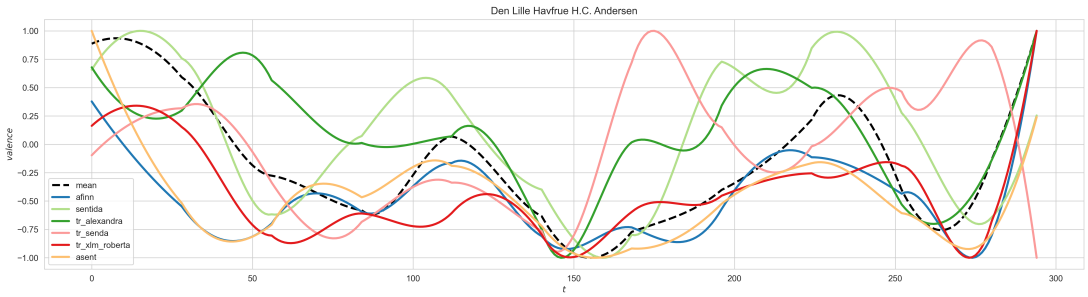


Figure 2: An example of visualized sentiment arcs of **The Little Mermaid**: Detrended arcs of systems and mean annotator score (black line). The x-axis represents the story progression in sentences.

Little Mermaid exhibits a particularly high correlation for detrended scores, with RoBERTa scoring 0.75 and Asent closely following at 0.72. While Asent's performance on *The Little Mermaid* is particularly surprising as its correlation with the raw arcs is close worse, we can see that transformers generally handle sentiment analysis for this task better than dictionary-based systems, both due to dictionary-based systems overemphasizing peaks in the sentiment arc (like Sentida in Fig. 2) or missing them (as Afinn, also Fig. 2). Transformers, however, appear to exhibit more extreme values, the distributions of their scores being less normal with a higher standard deviation than human and dictionary-based systems' scores (see the Appendix for the distribution of all scores).[23] In general, most models' performance improves when outlier effects are minimized.

[23]Note that the distribution of transformer scores may be an effect of using the confidence score for our transformation of their output labels (see section 2.4.4). Since the confidence scores of models tends to be relatively high (close to 0.9 in the range 0-1), using the confidence score for converting labels to values results in many high and low values.

3.3 Sentiment Analysis on Danish Religious Hymns

The analysis of Danish religious hymns presents a different pattern. Sentida consistently performs best among dictionary-based models in both original and modernized texts, achieving Spearman correlations of 0.49 and 0.53 respectively (Table 4). This suggests Sentida's rule-based approach, designed for short social media-like texts, captures the emotional tone in the hymns effectively.

Transformer-based models do not exhibit the significant advantage that they had in fairy tales. In the modernized hymns, RoBERTa shows a better correlation (0.46) than in the original (0.39), suggesting that modern language adaptations is more amenable to transformer processing, potentially due to the training data characteristics. But all transformer models perform worse than rule-based models. It is notable that the English systems, RoBERTa, VADER and Syuzhet, applied to Google translations, perform better than other systems. Syuzhet performs better than any other

	Afinn	Sentida	Asent	Alex.in.	Senda	RoB	VADER	Syuzhet	RoB
Hymns orig.	0.39	**0.49**	0.40	0.39	0.32	0.39	-	-	-
Hymns mod.	0.40	**0.53**	0.41	0.39	0.35	0.46	-	-	-
EN (baseline)	-	-	-	-	-	-	0.55	0.58	**0.66**

Table 4: Sentiment analysis of hymns: Spearman correlation between scores on the **original** (above) **modernized lines** (middle) and, for comparison, the three English system's scores of google-translated lines (below) to the human mean scores. Note that RoBERTa (RoB) on the right was used on translated sentences, and RoBERTa on the left on the original Danish sentences. Best performing Danish tools in bold, best baseline in green. Note that the Spearman's *rho* between the annotators of the hymns (n=2) is 0.726.

dictionary-based systems, possible due to it being developed for the literary domain.[24]

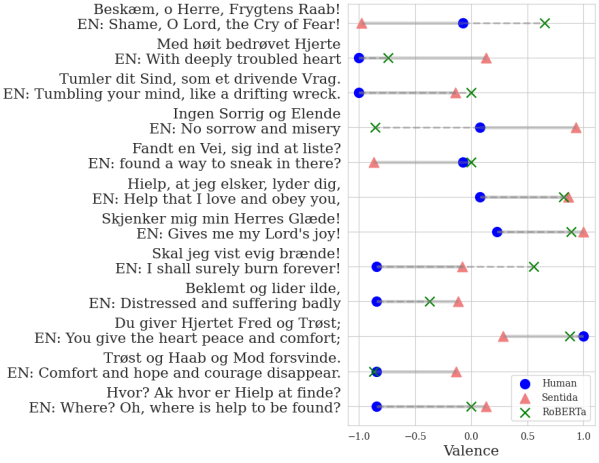


Figure 3: The 12 verses of the hymns with the highest absolute disagreement between human and Sentida score on original text (descending). RoBERTa scores are visualized for comparison. Validated English translation is supplied below the original Danish text.

An inspection of verses with the highest disagreement between human scores and scores of the best-performing Danish model (Sentida) of the original text suggests that disagreement results both from non-modern spelling and archaic vocabulary, but also from the genre and domain particularities of the hymns (Fig. 3). A clear example is the verse "beklemt og lider ilde": It contains both an overall archaic vocabulary and word-order, but models do not pick up on its negative tone. Even one archaic word in a verse appears to lead to errors for Sentida: In the line containing the archaic "hielp", humans rate the verse close to neutral, since the word suggests a wished-for state rather than an actuality, while models appear to weight the positive words in the verse highly, not observing the conditional. This is also the case for the first line, exhibiting the top disagreement, where the word "beskæm" (archaic) in combination with the poetic apostrophe ("o Lord") indicates the wish for God to "shame" in the sense of "reject" fear.

Apostrophes are not the only poetic feature that appears to confound model scores, generally, high-disagreement verses suggest that the genre and domain is a challenge to the models. For example, the last line in Fig. 3 employs repetition as well as the poetic exclamation "Ak", which may have prompted annotators to assign a very negative score, while models are blind sensitive to these genre- or domain-specific poetic devices. RoBERTa shows some similarities with Sentida, in this regard, with some overlap in which verses appear among the top disagreements with humans, like here the "Ak"-verse (for the top 12 verses with most disagreement of both models, see the Appendix).

3.4 Comparison

A comparison between the fairy tales and hymns reveals an essential reversal of fortune for the models taken into consideration. The fairy tales, which use language creatively in order to construe a relatively simple narrative, provide longer, richer sentences, and appear to allow transformers to leverage their ability to deal with complex syntactic and semantic interactions, leading to higher correlations especially in the detrended analyses. This aligns with what has been observed in several previous studies about the strength of transformers in handling varied and complex sentence structures and meanings (Li et al., 2023; Madusanka et al., 2023).

In contrast, the hymns are of poetic language, broken in short verses, often repetitive, figurative or allegoric, and heavily patterned. This kind of text seems to benefit less from the contextual capabilities of transformers. The short nature of

[24] Although domain-specific tools tend to rely on less data, the Syuzhet dictionary is relatively large: developed from 165,000 human-coded sentences from contemporary literary novels in the Nebraska Literary Lab (Jockers, 2015).

the verses, the weight of single words (compared to their weight in more complex interactions of narrative prose), and poetic devices, seems to allow dictionary and rule-based methods to shine, while they might be reducing the effectiveness of Transformer-based sentiment analysis: not only do the dictionary-based models' go up, but the Transformers' performance go down, compared to the correlation in the fairy tales.

The consistently high performance of Sentida across different types of texts suggests that some rule-based systems, especially those tailored or adapted to specific languages like Danish, can effectively capture sentiment even without the contextual depth provided by transformers, especially where historical language is being treated. Still, both the baseline models, VADER and Syuzhet applied to google-translated text, also show a good performance – and consistently so – outperform Danish models in the Hymns, while constituting a robust alternative for the fairy tales as well.

3.4.1 Comparing the two best-performing Danish models

As Sentida and RoBERTa were the best-performing systems, we computed the word-level SHAP-values from RoBERTa's output (applied to Danish) to compare them to the weights indexed in Sentida.[25] SHAP-values are used to understand models' predictions, gauging the importance of individual features in informing the predicted label (in the case of RoBERTa, the role of individual tokens in positive, neutral, or negative results)(Lundberg and Lee, 2017). The process involves calculating the contribution of each word by removing it and observing the change in the model's prediction. The impact of context (preceding/following words) is addressed by iterating this process over permutations of the words.[26]

As can be seen in Table 5, RoBERTa's word-level SHAP scores explain a higher proportion of the variance in Sentida scores for Andersen's fairy tales compared to Hymns, both for positive and negative sentiments.The model's ability to predict positive sentiment variance is slightly stronger in the fairy tales (0.20) than in the hymns (0.16). The same pattern holds for negative sentiment, though the difference is less pronounced (0.15 vs. 0.13)

	POS (H)	NEG (H)	POS (A)	NEG (A)
R^2	0.16	0.13	0.20	0.15

Table 5: The R^2 score of regression models on Sentida's scores of words and SHAP-score (viz. RoBERTa word-weights) for Hymns (H) and Andersen's fairy tales (A) – i.e., the R^2 score represents the percentage of the variance in Sentida scores that SHAP-scores explain, ranging between 0 (no explanation), to 1 (complete explanation of variance).

The difference in R^2 between the two datasets suggests both vocabulary differences and that RoBERTa is actually acting "more like" Sentida on the data it performs best on (HCA). Andersen might use more frequent words and/or words that are simpler semantically and thus easier to agree upon across these two different systems. Moreover, as is also visualized in Fig. 4, more than half of the words in the fairy tales (55.6%) and close to half in the hymns (48.9%) are assigned a 0 score by Sentida, while RoBERTa tends to assign more words a positive or negative value (see the Appendix for a list of top-positive and negative words not recognized by Sentida).

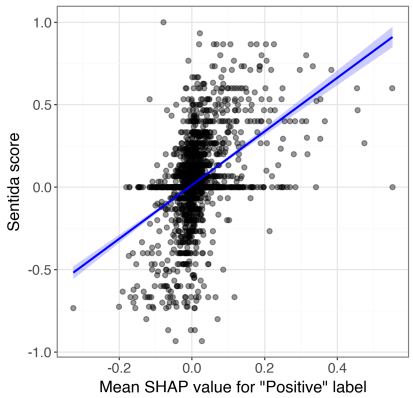


Figure 4: Visualization of the correlation – in the case of fairy tales (HCA) – between the Sentida score of words (y-axis) and their corresponding SHAP-score of the RoBERTa model (x-axis), here, the degree to which the word contributes to the model assigning the "positive" label.

Considering that Roberta (on Danish) underperforms with respect to Sentida on the hymns, which are evaluated at the verse level, discrepancies in their vocabularies can be illuminating. Given the reduced dimension of poetic verses, sentimental evaluation at that level has less to do with syn-

[25] Note that we used a custom tokenization: instead of using the RoBERTa tokenization, which usually splits one Danish word into multiple tokens, we consider one word (whitespace separated) as one token for the SHAP analysis.

[26] In our case, 10 random forward and backward permutations (20 in total), after which we average the differences between SHAP-values of permutated features and original features, as implemented in the SHAP Python package: PermutationExplainer

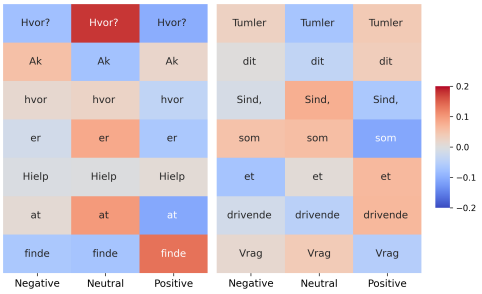


Figure 5: Two sentence examples (left, right). The heatmap shows how words contribute to the label (Negative, Neutral, Positive) assigned to sentences by RoBERTa, where weights are measured as SHAP-scores. Higher values (in red) signify how much a word contributes *toward* the label (on the x-axis) and the bluer the word, the more it contributes *away* from the label assigned. For example, 'finde' in the sentence on the left contributes toward the positive label. The sentences are ordered from top to bottom.

tax or larger discourse-structures, and much more to do with the interplay of individual words' nuances. The nature of the poetic language often adopted in the hymns, that tends to weigh on the contrast and association of terms, might also give a particularly important role to lexical semantics in the overall valence of each verse. It is not too far-fetched to imagine that the scores of Sentida, manually curated and directly assigned by annotators (n=3)(Lauridsen et al., 2019), are the real point of advantage of the model in these circumstances.

As salient examples, we also examined the SHAP-scores of two sentences, which both occur among sentences with top disagreement between human scores and the RoBERTa and Sentida model (see sentences in appendix). While humans rated a sentence like the leftmost of Fig. 5 very negatively (see Fig. 3), we find RoBERTa labelling it neutral, mostly due to the words 'finde' and 'Hvor?' pulling it in opposite directions. Notably, the model does seem to recognize a difference between 'Hvor' with and without the questionmark, and does recognize the poetic exclamation ('Ak') as somewhat negative, suggesting a sensitivity to the register. As suggested before, it appears not to process the 19th century spelling of 'hielp' (hjælp) adequately, which has close to a 0 SHAP-score for all labels. Similarly, for the rightmost sentence in Fig. 5, the negative poetic imagery which may make humans rate the sentence accordingly is not reflected in the SHAP-scores of words, notably with the negatively associated "vrag" (wreck) weighted toward netural.

4 Conclusion and Future Works

We have tested sentiment analysis tools on Danish literary prose and poetry, using a small collection of historical fairy tales by H.C. Andersen for prose and of traditional religious hymns for poetry. Our goal was to study the abilities and limitations of SA methodologies in handling a particularly low-resource setting: relatively low-resource language on low-resource domains. Employing both human annotators and a range of sentiment analysis models, we have shown that transformer-based models generally outperform dictionary-based systems in the analysis of fairy tales, especially when considering detrended scores – consistent with previous work (Bizzoni and Feldkamp, 2023; Allaith et al., 2023; Schmidt et al., 2021). These models seem to have a better ability to interpret the emotional and narrative structures of fairytales more effectively, and better mimic the human experience of reading narrative fiction. However, for the poetic hymns with short verses, the performance gap between transformers and dictionary-based models changes, and dictionary-based approaches, especially Sentida, show better performance. A combination of approaches may be explored in the future, as our comparison using SHAP-scores suggests that models capture different aspects of texts. Including more texts from different authors in the dataset may also give a more nuanced picture of SA in Danish, and it should be noted that the prose part of our corpus – a single author – may bias the results. Still, as Danish resources are consistently outperformed, both by the multilingual model or by the English baseline models applied to raw Google translations, we observe that there is a need for developing a Danish-based model for SA of literary texts across genres and periods.

In future, we would like to expand the dataset to include a broader range of genres and apply more models and model adaptions. Integrating comprehensive historical, semantic, and emotional lexica, may also improve the granularity and accuracy of sentiment predictions. Further refining detrending techniques may also be beneficial, particularly for texts where narrative context heavily influences sentiment interpretation. Finally, more extensive collaboration between linguists and literary scholars may help refine the algorithms used, embedding deeper literary and linguistic insights into the development of sentiment analysis tools for treating specific language use of the literary domain.

Limitations

We want to underline that our results are based on a limited set of Danish literary historical texts and should be interpreted accordingly. It should also be noted that the prose part of our corpus – consisting of a single author (whereas hymns have several authors) – may bias our results. Moreover, the demographic of our dataset is reduced (in terms of gender, ethnicity, age, social class, etc.). While this work has aimed to test Danish resources for continuous sentiment analysis, there are various other English-based resources which may perform better than the ones selected here – especially more recent generative methods.

References

Huwail J. Alantari, Imran S. Currim, Yiting Deng, and Sameer Singh. 2022. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, 39(1):1–19.

Ali Allaith, Kirstine Degn, Alexander Conroy, Bolette Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment Classification of Historical Danish and Norwegian Literary Texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional Sequencing and Development in Fairy Tales. In *Affective Computing and Intelligent Interaction*, pages 668–674, Berlin, Heidelberg. Springer.

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the Future with Social Media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2020. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLoS ONE*, 15(11).

Yuri Bizzoni and Pascale Feldkamp. 2023. Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study. In *Proceedings of the 3rd International Workshop on Natural Language Processing for Digital Humanities*, pages 219–226, Tokyo, Japan. Association for Computational Linguistics.

Yuri Bizzoni and Pascale Feldkamp. 2024. Below the sea (with the sharks): Probing textual features of implicit sentiment in a literary case-study. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 54–61, Malta. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022a. Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of Andersen's fairy tales. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022b. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Wayne C. Booth. 1983. *The Rhetoric of Fiction*, 2nd edition edition. University of Chicago Press, Chicago.

Katherine Bowers and Quinn Dombrowski. 2021. Katia and the Sentiment Snobs. Blog: Datasitter's Club.

Center for Children's Literature and Media CCLM. 2003. Danske børn og unge har stort kendskab til H.C. Andersen.

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.

Luna De Bruyne, Pranaydeep Singh, Orphee De Clercq, Els Lefever, and Veronique Hoste. 2022. How language-dependent is emotion detection? evidence from multilingual BERT. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

Katherine Elkins and Jon Chun. 2019. Can Sentiment Analysis Reveal Structure in a Plotless Novel? ArXiv:1910.01441 [cs].

Hady Elsahar and Matthias Gallé. 2019. To Annotate or Not? Predicting Performance Drop under Domain Shift. 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 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in *Never Let Me Go* : Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Roman Jakobson. 2010 (1981). Linguistics and poetics. In *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton.

Jianbo Gao, H. Sultan, Jing Hu, and Wen-Wen Tung. 2010. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *IEEE Signal Processing Letters*, 17(3):237–240.

Matthew Jockers. 2014. A Novel Method for Detecting Plot. Matthew L. Jockers Blog.

Matthew L. Jockers. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

Gustav Aarup Lauridsen, Jacob Aarup Dalsgaard, and Lars Kjartan Bacher Svendsen. 2019. SENTIDA: A New Tool for Sentiment Analysis in Danish. *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift*, 4(1):38–53. Number: 1.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023. Assessing the Capacity of Transformer to Abstract Syntactic Representations: A Contrastive Analysis Based on Long-distance Agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33. Place: Cambridge, MA Publisher: MIT Press.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tom Lundskær-Nielsen. 2014. The Language of Hans Christian Andersen's Fairy Tales – Compared with Earlier Tales. *Scandinavistica Vilnensis*, 1(9):97–112. Number: 9.

Tharindu Madusanka, Riza Batista-navarro, and Ian Pratt-hartmann. 2023. Identifying the limits of transformers when performing model-checking with natural language. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3539–3550, Dubrovnik, Croatia. Association for Computational Linguistics.

Marita A. Nielsen. 2020. Salmesprog. In *Dansk Sproghistorie Bind 4. Sprog i brug*. Aarhus University Press and Society for Danish Language and Literature (DSLDK).

Bruno Ohana, Sarah Jane Delany, and Brendan Tierney. 2012. A Case-Based Approach to Cross Domain Sentiment Classification. In *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, pages 284–296, Berlin, Heidelberg. Springer.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrew J. Reagan, Christopher M. Danforth, Brian Tivnan, Jake Ryland Williams, and Peter Sheridan Dodds. 2017. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6(1):1–21. Number: 1 Publisher: SpringerOpen.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ Data Science*, 5(1):1–12.

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2).

Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A. Vouros. 2009. Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria. Association for Computational Linguistics.

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Brian Richardson. 2000. Linearity and Its Discontents: Rethinking Narrative Form and Ideological Valence. *College English*, 62(6):685–695.

Dan Ringgaard and Mads Rosendahl Thomsen, editors. 2017. *Danish literature as world literature*. Literatures as world literature. Bloomsbury Academic, New York.

Louise M. Rosenblatt. 1982. The literary transaction: Evocation and response. *Theory Into Practice*, 21(4):268–277.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Schmidt and Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. *Fabrikation von Erkenntnis: Experimente in den Digital Humanities -*.

Nina Schneidermann and Bolette Pedersen. 2022. Evaluating a New Danish Sentiment Resource: the Danish Sentiment Lexicon, DSL. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pages 19–24, Marseille, France. European Language Resources Association.

Morten Skovsted, Mads Djernes, Kirsten Nielsen, Martin Horsntrup, and Hanne J. Jakobsen. 2019. Hvad gør en ny salme til en god salme? *Salmedatabasen*.

Elias Stengel-Eskin, Jimena Guallar-Blasco, and Benjamin Van Durme. 2021. Human-model divergence in the handling of vagueness. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 43–57, Online. Association for Computational Linguistics.

Hsiu-Yuan Tsao, Ming-Yi Chen, Hao-Chiang Koong Lin, and Yu-Chun Ma. 2018. The asymmetric effect of review valence on numerical rating: A viewpoint from a sentiment analysis of users of TripAdvisor. *Online Information Review*, 43(2):283–300. Publisher: Emerald Publishing Limited.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

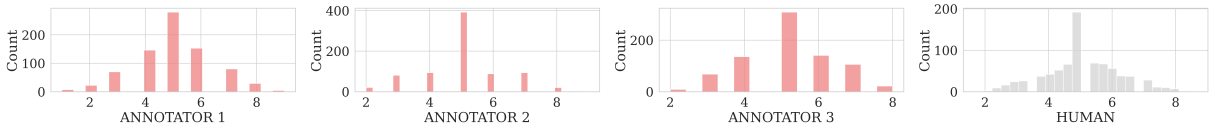
Yuemei Xu, Han Cao, Wanze Du, and Wenqing Wang. 2022. A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations. *Data Science and Engineering*, 7(3):279–299.

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of Happy Endings in German Novels Based on Sentiment Information. In *Interactions between Data Mining and Natural Language Processing*, pages 9–16, Riva del Garda.

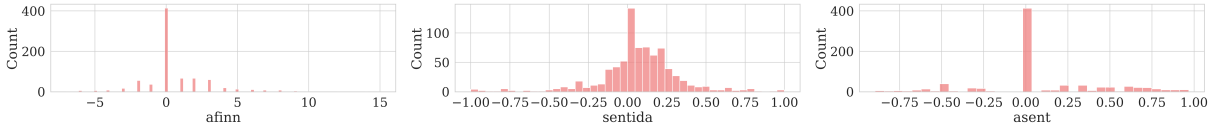
A Appendix

Verse	English translation	Human	RoB	Sentida
Beskæm, o Herre, Frygtens Raab!	*Shame, O Lord, the Cry of Fear!*	-0.08	0.65	-0.98
Med høit bedrøvet Hjerte	*With deeply troubled heart*	-1.00	-0.74	0.13
Tumler dit Sind, som et drivende Vrag.	*Tumbling, your mind, like a drifting wreck.*	-1.00	0.00	-0.14
Ingen Sorrig og Elende	*No sorrow and misery*	0.08	-0.86	0.93
Fandt en Vei, sig ind at liste?	*Found a way to sneak in there?*	-0.08	0.00	-0.87
Hielp, at jeg elsker, lyder dig,	*Help that I love and obey you,*	0.08	0.82	0.87
Skjenker mig min Herres Glæde!	*Gives me my Lord's joy!*	0.23	0.89	1.00
Skal jeg vist evig brænde!	*I shall surely burn forever!*	-0.85	0.55	-0.08
Beklemt og lider ilde,	*Distressed and suffering badly*	-0.85	-0.37	-0.12
Du giver Hjertet Fred og Trøst;	*You give the heart peace and comfort;*	1.00	0.88	0.28
Trøst og Haab og Mod forsvinde.	*Comfort and hope and courage disappear.*	-0.85	-0.87	-0.13
Hvor? Ak hvor er Hielp at finde?	*Where? Oh, where is help to be found?*	-0.85	0.00	0.13
Til Smerte, Spot og Spe!	*For pain, ridicule, and mockery!*	-1.00	0.00	-0.55
Tumler dit Sind, som et drivende Vrag.	*Tumbling, your mind, like a drifting wreck.*	-1.00	0.00	-0.14
Skjuler mig for Synd og Død,	*Hides me from sin and death,*	-0.08	-0.93	-0.49
Af Pine, Kval og Plage	*Of torment, anguish, and suffering*	-0.85	0.00	-0.38
Fra Forkrænkelse og Død;	*From Violation and Death;*	-0.85	0.00	-0.73
Ei Trøst jeg fandt, ei Lindring kom	*No comfort I found, no relief came*	-0.85	0.00	-0.30
Hvor? Ak hvor er Hielp at finde?	*Where? Oh, where is help to be found?*	-0.85	0.00	0.13
Paa Jorden er der Strid og Had,	*On Earth there is strife and hate,*	-0.85	0.00	-0.37
Mishaab og Strid har hver timelig Stund.	*Hopelessness and strife have each earthly hour.*	-0.85	0.00	-0.29
Ham det fryder, at Dødsstriden	*He delights that the struggle of death*	0.08	-0.92	0.47
Gjør dit Guld Dig frydefuld?	*Does your gold make you joyful?*	-0.08	-0.92	0.53
Forkast da Barnet ei, som kommer	*Do not reject the child who comes*	-0.08	-0.92	-0.09

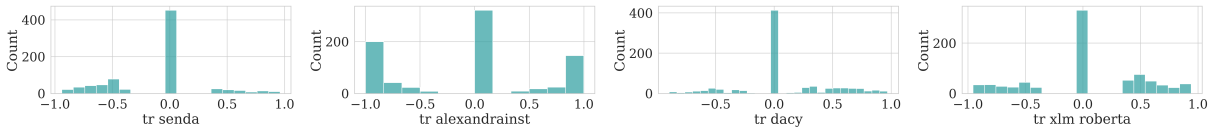
Table 6: The 12 verses of the **Hymns** that exhibit the highest absolute disagreement between the human mean vs Sentida (top 12 rows) and vs RoBERTa (RoB) scores (bottom 12). Highlighted rows recur in the top 12 disagreement-verses of both RoBERTa and Sentida. Note that human mean values tend to recur: due to two annotators for the hymns, only whole and half numbers within the 0-10 range are possible, so that normalized values reflect this.



(a) Distribution of annotators' and the human mean scores.

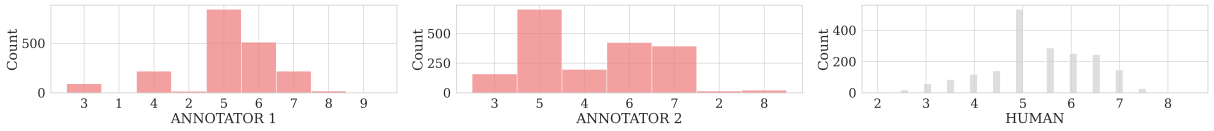


(b) Distribution of dictionary-based methods' scores.

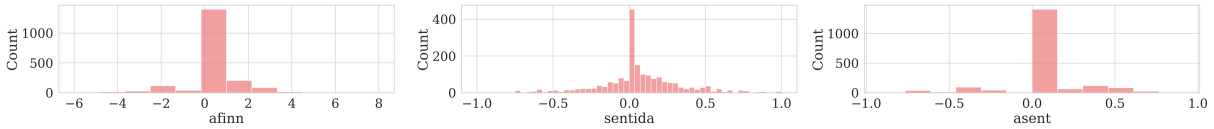


(c) Distribution of transformer-based methods' scores.

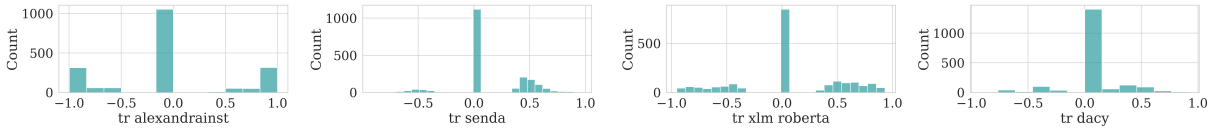
Figure 6: **HCA dataset**, distributions of scores per Sentiment Analysis method.



(a) Distribution of annotators' and the human mean scores.



(b) Distribution of dictionary-based methods' scores.



(c) Distribution of transformer-based methods' scores.

Figure 7: **Hymns dataset**, distributions of scores per Sentiment Analysis method.

Word	**English translation**
mærkværdig	*odd*
fornemste	*most distinguished*
fineste	*nicest*
underligt	*curious/strange*
klogeste	*smartest*
herligt	*magnificent*
underlig	*curious/strange*
klogt	*smart*
pragt	*splendor*
morsomt	*funny*
nedrig	*lowly*
styg	*hideous*
skammede	*shamed*
nykker	*whims*
kostbart	*precious*
kalkunkylling	*turkey chicken*
fangst	*catch*
værre	*worse*
forvildet	*bewildered/lost*
grueligste	*most gruesome*

Table 7: Top 10 positively (top) and negatively (bottom) weighed words of RoBERTa (as gauged via SHAP-scores) in the fairy tales that are not indexed in Sentida (the stem of some words, like "klog", are indexed in Sentida, yet it is unrecognized with the 't'-ending). Note that while most words appear to be reasonably justified for their positive/negative label, some artifacts appear (e.g. "kalkunkylling").