Continuous sentiment scores for literary and multilingual contexts

Laurits Lyngbaek[1], Pascale Feldkamp[1], Yuri Bizzoni[1], Kristoffer L. Nielbo[1], and Kenneth Enevoldsen[1]

[1] Center for Humanities Computing, Aarhus University, Aarhus, Denmark

Abstract

Sentiment Analysis is widely used to quantify sentiment in text, but its application to literary texts poses unique challenges due to figurative language, stylistic ambiguity, as well as sentiment evocation strategies. Traditional dictionary-based tools tend to underperform, especially for low-resource languages, and transformer models, while promising, output coarse categorical labels that limit fine-grained analysis. We introduce a novel continuous sentiment scoring method based on concept vector projection, trained on multilingual literary data, which captures nuanced sentiment expressions across genres, languages, and historical periods. Our approach outperforms existing tools on English and Danish texts, producing sentiment scores which distribution matches human ratings, improving sentiment arc modeling and analysis in literature.

Keywords: sentiment analysis, computational literary studies, historical texts, semantic embeddings

1 Introduction & Related Works

Sentiment analysis quantifies sentiment in text and is widespread across domains, from product reviews analysis to social media monitoring [7; 32]. Computational literary studies have employed sentiment analysis to model narrative dynamics, particularly sentiment arcs, across novels [5; 16; 26; 38]. This requires continuous sentiment scores, mapping sentiment along a spectrum rather than using categorical labels like positive/negative.

Despite the growing use of continuous sentiment scoring in literary studies, the validity of current tools in capturing literary sentiment expression remains underexplored. Popular tools such as *Syuzhet* have faced severe criticism for oversimplification or poor generalizability [31] – issues that point to broader limitations in applying off-the-shelf sentiment tools to literary texts. Indeed, the literary domain poses distinct challenges: figurative language, multiple narrative layers, and stylistic ambiguity all complicate sentiment detection [3; 8].

More recent transformer-based models appear better equipped to handle the complexity of literary language [30], and techniques exist to transform categorical model outputs into continuous scores [4]. This method has proven more effective than tailored dictionary-based tools, particularly in low-resource language settings and across languages [13]. However, empirical benchmarks comparing model predictions to human judgments remain limited in languages other than English.

We identify three main issues where current methods see a noticeable performance drop:
1) **Cross-lingual** performance drops. Most Sentiment Analysis tools tackle high-resource languages, and their transfer to low-resource ones like Danish is non-trivial. Although Danish has

Laurits Lyngbaek, Pascale Feldkamp, Yuri Bizzoni, Kristoffer L. Nielbo, and Kenneth Enevoldsen. "Continuous sentiment scores for literary and multilingual contexts." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 447–464. https://doi.org/10.63744/nVu1Zq5gRkuD.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

several dictionary-based tools (i.a., Asent [12], Sentida [21]), these have seen little use on historical literature and struggle with complex literary forms. Comparing tools for Danish literary sentiment analysis, Feldkamp et al. [13] found that multilingual transformer models outperfromed dictionaries – likely because they leverage contextual attention. While multilingual transformers, such as mBERT and XLM-R [11], show promise for cross-lingual sentiment analysis in literature [13], cultural and linguistic biases inherited from English pretraining remain a concern [10; 37].
2) **Cross-domain** performance often drops, especially when applying tools trained on social media to literature, where sentiment is expressed in a distinct and complex manner [3; 14; 33]. Literary language tends to be more *omissive and implicit*, relying less on charged vocabulary and more on *concrete* descriptions of objects and situations to evoke affect – a domain-specific mode of sentiment expression that models fail to capture [14]. This domain-specificity varies across domains: when using a model fine-tuned on Twitter posts, poetry shows the weakest correlation with human ratings, prose falls in the middle, and Facebook posts show the strongest correlation [14].
3) **Historical data**, marked by diachronic language change, reduces model performance. While fine-tuned multilingual transformers show promise [1; 13; 29], challenges remain. Lexical drift – including semantic shift, word loss (e.g., *thou*, *peradventure*), changing frequencies, and temporal polarity shifts – limits sentiment inference if models rely on priors from modern corpora.[1] For temporal polarity shifts, even short-term changes can lower model performance [23].

A major drawback of recent transformer-based approaches is that, while they outperform dictionary-based tools on historical and literary data [13], they tend to perform trinary classifications (positive, neutral, negative), limiting their usefulness for fine-grained sentiment analysis. Although model confidence scores can be repurposed for continuous output – with medium to strong correlation to human ratings [13] – the resulting distributions still cluster around the original three categories, producing what is effectively a pseudo-trinary distribution. This poses a problem for literary analysis tasks, not least sentiment arc modelling, where detrending methods to smoothen out the signal necessitate continuous scores. When sentiment scores behave in extreme ways – as they will with pseudo-trinary distributions – smoothing will tend to collapse variation toward the neutral midpoint, removing meaningful information.

In this paper, we introduce a method to create continuous-scale sentiment scores that are more closely aligned with the distribution of human scores, while also mitigating language-, domain-, and historical data issues by basing the method on the language and domain of the use case.

We test this approach on English and Danish literary texts, comparing it to existing transformer-based models and popular dictionary-based tools, across both fiction and nonfiction genres. The benchmark includes both historical literary genres (e.g., 19th-century hymns) and contemporary texts (e.g., blogs), enabling us to evaluate model performance in settings that better reflect the needs of researchers working with multilingual or diachronic literary corpora. We pursue three aims: (1) to assess model performance on contemporary literary and non-literary texts; (2) to compare performance across literary genres; and (3) to evaluate models on historical and multilingual literary data. We begin by testing our approach on *Fiction4* — a recent annotated fiction corpus that spans four literary genres, two languages (English and Danish), in the period 1798 to 1965. We then validate our approach further on *EmoBank*, a standard sentiment analysis dataset that includes contemporary genres and a small set of fiction, to gauge the generalizability of our approach and to control for overfitting to literary data.

[1] Diachronic sentiment analysis is challenging for traditional machine learning approaches as words' meaning and polarity change in a continuous way, while most models require steady ground truths for training, creating artificial "museums" of words' sentiment scores in a given historical period.

2 Methods

2.1 Data

Dataset	Period	N annotations	N words	$\bar{x}$ words/sentence	N annotators
↓ *EmoBank*	1990-2008	8,870	143,499	16.18	10
Letters		1,413	21,639	15.31	10
Blog		1,336	20,874	15.62	10
Newspaper		1,314	25,992	19.78	10
Essays		1,135	26,349	23.21	10
Fiction		2,753	31,491	11.44	10
Travel-guides		919	17,154	18.67	10
↓ *Fiction4*	1798-1965	6,300	73,250	11.6	>2
Hymns	1798-1873	2,026	12,798	6.3	2
Fairy tales	1837-1847	772	18,597	24.1	3
Prose	1952	1,923	30,279	15.7	2
Poetry	1965	1,579	11,576	7.3	3

Table 1: Datasets with valence annotation. Valence was annotated on a sentence basis, so 'N annotations' indicates the number of sentences. The total number of sentences considered is $n = 15,170$. 'N annotators' indicates the number of annotators reported per sentence.

Emobank is a text corpus manually annotated for sentiment according to the psychological Valence-Arousal-Dominance scheme. It was compiled at *JULIE Lab*, Jena University [9],[2], containing sentences from the MASC dataset, which is diverse both in terms of overall composition with diverse domains, and topically within categories.[3] It includes six categories: Letters, Blog, Newspaper, Essays, Fiction, and Travel guides.[4] Inter Rater Reliability for the whole dataset is: Krippendorff's $\alpha = 0.34$.[5] We use the mean sentence-based valence scores overall and per category to compare model performance.

Fiction4 is a dataset of literary texts, spanning literary texts across four genres and two languages (English and Danish) in the 19th and 20th century.[6], compiled at the *Center for Humanities Computing*, Aarhus University. The corpus consists of three main authors, Sylvia Plath for poetry, Ernest Hemingway for prose, and H.C. Andersen for fairytales. Hymns were collected from Danish official church hymnbooks published between 1798 and 1873. All sentences in the corpus were annotated for by at least two annotators [14]. Inter Rater Reliability for the whole dataset is: Spearman's $\rho = 0.63$ and Krippendorff's $\alpha = 0.67$.[7] We use the mean sentence-based valence score overall, per language set, and per genre to compare model performance.

[2] `https://github.com/JULIELab/EmoBank/`

[3] On some *EmoBank* categories: *Essays* includes eight texts, i.a., "A Brief History of Steel in Northeastern Ohio'. *Fiction* comprises six prose pieces across genres, i.a., Richard Harding's "A Wasted Day" and the SciFi story "Captured Moments". *Newspapers* contain reports (e.g., "A.L. Williams Corp. was merged into Primerica Corp.") and longer reportages. *Travel Guides* are written in prose, including both place histories (e.g., "A Brief History of Jerusalem") and reflective pieces (e.g., "Dublin and the Dubliners"). See the full MASC corpus at: `https://anc.org/data/masc/corpus/browse-masc-data/`.

[4] We excluded the 'Sem-Eval' category as it was internally diverse.

[5] Since *EmoBank* lacks unique annotator IDs, we cannot correlate individual annotators' scores. Instead, Krippendorff's α measures agreement across ratings per item. IRR per subset is shown in Table 4.

[6] `https://huggingface.co/datasets/chcaa/fiction4sentiment`, for details, see [14]

[7] Humans rarely reach an agreement higher than 80% (α>0.80) for categorical tagging (positive/neutral/negative) on *nonliterary texts* [36] – and have lower IRR for continuous scale annotation [2] – especially of literary texts [27].

2.2 Comparison models

2.2.1 Dictionary-based

Because of their popularity and wide usage in literary studies [1; 4; 6], as a baseline, we tested the dictionary-based tools `VADER` [15] and `Syuzhet` [17]. They assign sentiment scores (from negative to positive) by word-score matching and specific rules. `Syuzhet` was developed explicitly for literary texts.[8] When using these tools, we translated Danish sentences into English as they do not perform well on the original Danish.[9] As such, dictionaries represent a rough baseline.

2.2.2 Transformer-based

To test transformer-based methods, we chose two multilingual models. When testing models on Danish texts, we added three models fine-tuned for Danish. These were all tested across *EmoBank* categories, as well as *Fiction4* genres and languages. We list all models in Appendix A, Table 5.[10] One of the multilingual models – `twitter-xlm` – showed the best performance on *Fiction4* in Feldkamp et al. [13]. Danish models were picked based on their performance in a recent benchmark [20], and – in the case of `MeMo-BERT-Sa` – for being developed for 19$^{\text{th}}$-century novels [19].

Conversion of model output: We convert Transformers' standard three-ways outputs (positive, neutral, negative) to continuous values using their confidence scores [11] as a proxy for intensity (e.g., *positive*, 0.67 → +0.67; *negative* → –0.67; *neutral* → 0). Mapping a model's confidence values to a continuous scale often outperforms dictionary-based tools for literary sentiment [4; 13].

$$\text{intensity} = \begin{cases} +p, & \text{if positive,} \\ 0, & \text{if neutral,} \\ -p, & \text{if negative.} \end{cases}$$

We tested transformer-based models on the original Danish and English, as well as on the Danish sentences translated to English (see Table 3), since one study found that some models work better on google-translated sentences [13], perhaps as the translation acts to standardize historic forms.

2.3 Our approach

It has been claimed that concepts – such as a sentiment – are approximately represented in a linear fashion within embedding space, which is denoted by the linear representation hypothesis [24]. The hypothesis states that concepts are encoded as a direction in the embedding space and that the further you move in a given direction, the stronger the concept is represented (see Figure 1). These linear representations of semantic information have been found in both encoding and decoding models, at varying levels of abstraction [22; 34; 35; 39]. Suppose we have access to the direction that encodes sentiment. In that case, we can project any embedded sentence onto the concept vector and gauge the sentiment of any given sentence, as seen in Figure 1.

2.3.1 Concept Vector Projection

We propose an algorithm that constructs a concept vector in a given embedding space using positive and negative exemplary sentences that represent the opposing extremes of the concept. Using

[8] The `Syuzhet` lexicon was developed in the *Nebraska Literary Lab* under the direction of Matthew L. Jockers.

[9] Using `googletrans`: `https://pypi.org/project/googletrans/`. Humans did not review translations.

[10] Code for comparing (HuggingFace-stored) sentiment models (with transformed outputs) on the *Fiction4* or *EmoBank* is at: `https://github.com/centre-for-humanities-computing/literary_sentiment_benchmarking`.

[11] The score output by finetuned models (e.g., "positive", 0.66) is a softmax-normalized class probability – a pseudo confidence score – reflecting how strongly a model prefers one label over another. It comes from the linear classification head atop models.

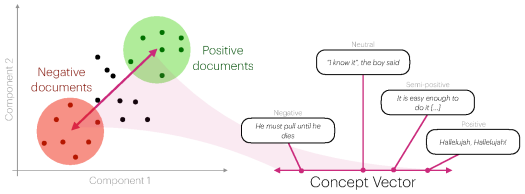

Figure 1: An overview of how a concept vector for sentiment is constructed and what information it contains. A circle represents an embedded document.

a pre-trained sentence embedding model **M**, the algorithm embeds a labeled set of sentences **S**. It assumes that a concept – here sentiment – is represented linearly in the embedding space. To define the concept vector, the algorithm computes the mean embedding of both the positive and negative sentiment examples, then calculates the vector pointing from the negative to the positive mean. This relies on the assumption that when averaging multiple sentences, all non-sentiment information will disappear as Gaussian noise with a mean of zero, leaving behind only the information encoding sentiment [18; 39].

The resulting vector then theoretically encodes sentiment direction. New sentences can be assessed for their relation to the sentiment by projecting their embeddings onto this vector: the farther along the direction the projection lies, the stronger their positive relation is. Defining the concept vector as a unit vector, the projection of a given embedding $\mathbf{e}_i$ onto the unit concept vector $\hat{\mathbf{v}}$ is given by the dot product: $\mathbf{e}_i \cdot \hat{\mathbf{v}}$. This projects the sentence embedding to the subspace spanned by the Concept Vector. The high-dimensional embedding has thereby been reduced to a one-dimensional sentiment score, as seen in figure Figure 2. Defining a concept vector requires only a set of positive and negative example sentences. This suffices to predict the sentiment of any subsequent sentence, whether labeled or unlabeled. The Concept Vector Projection (CVP) algorithm formally described in Appendix C. The implementation of this method is available at `https://github.com/centre-for-humanities-computing/embedding-projection`.

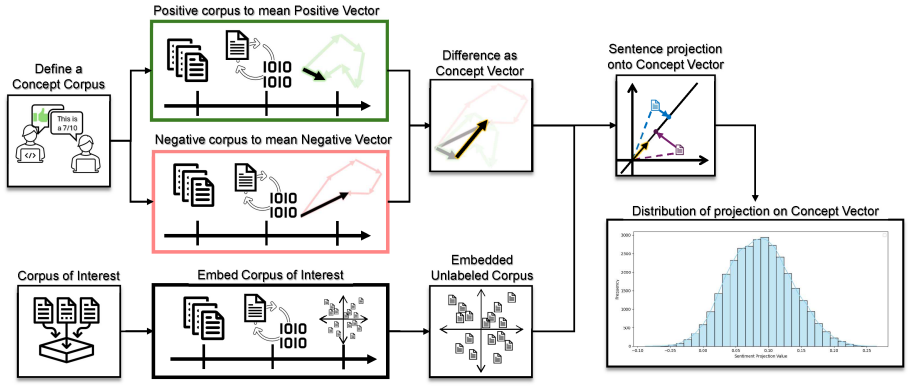

Figure 2: A visualization of how the Concept Vector Projection is constructed. It shows how to use a labeled sentiment corpus to predict sentiments of an unlabeled corpus of interest. The vectors shown are reduced to a two-dimensional Euclidean space for visualization, but normally reside in a high-dimensional space.

2.4 Models

The implementation of Concept Vector Projection used to classify sentiment in this paper is based on the language model `paraphrase-multilingual-mpnet-base-V2`[12] [28]. This is a 278M parameter model, based on a mean-pooled BERT architecture, optimized for sentence similarity by using Siamese and Triplet networks. This model was chosen because of its multilingual capabilities and excellent size-to-performance ratio. Investigations during model selection indicate that a larger model may increase model correlation with human ratings in exchange for compute budget.

Our Concept Vector was defined using a training dataset of sentences with positive and negative sentiments from the *Fiction4* dataset. Since the sentences were originally rated on a numerical scale (1-9), they were translated to positive/negative ratings for the algorithm. We converted the mean ratings into ordinal labels through preset thresholds. That is, for the *Fiction4* ratings, we define:

$$\text{label} = \begin{cases} \text{positive}^{+} & \text{if rating} \geq 7 \\ \text{neutral}^{\varnothing} & \text{if } 7 > \text{rating} > 3] \\ \text{negative}^{-} & \text{if rating} \leq 3 \end{cases}$$

All the neutral sentences and 60% of the positive and negative sentences were in the *Fiction4* testing set. The remaining 40% were in a Concept Corpus of 204 positive and 168 negative sentences used to define the model's concept vector.

3 Results

3.1 Continuous scoring

A key benefit of the Sentiment Projection model is its ability – like dictionary tools – to produce genuinely continuous predictions. In contrast, Transformer-based token-classification models such as `xlm-roberta`, which can be coerced to output continuous scores (see subsubsection 2.2.2), in practice exhibit a "pseudo-trinary" behavior: their predictions cluster heavily at zero and at the two polar extremes. This behavior is visible both in the scatterplots of predicted vs true sentiments (Figure 3) and in the histograms of model outputs (Appendix A, Figure 5). When looking at the *EmoBank* results (Appendix A, Figure 4), the discretized output of `xlm-roberta` appears even more sharply tri-modal than the human scores, which average ten annotators.

[12] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

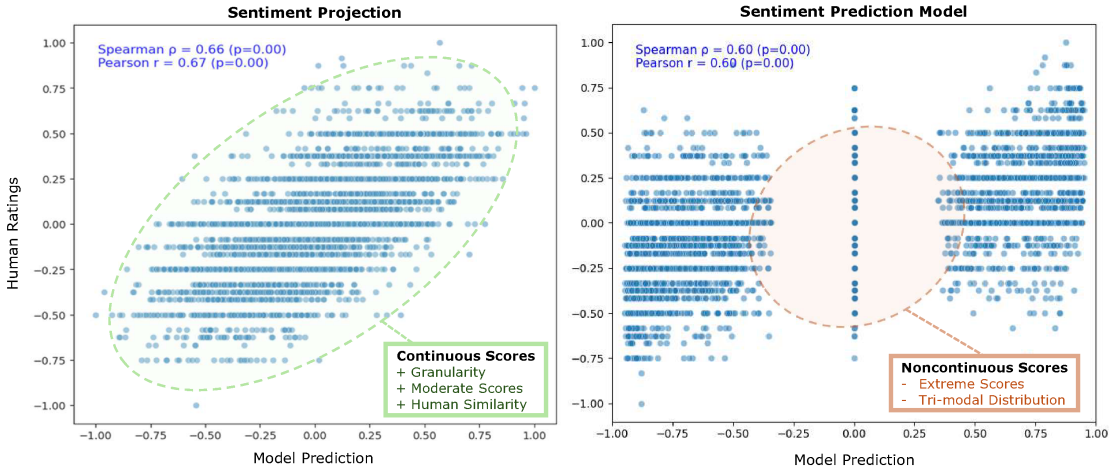

Figure 3: Scatterplot of Sentiment Predictions for respectively `Sentiment Projection` and `xlm-roberta`. While the `xlm-roberta` model, in theory, can predict a continuous space of sentiments when transforming it with confidence scores, inspection shows that certain ranges of the sentiments spectrum are not used. While both models achieve high correlations, it appears that `xlm-roberta` achieves this by matching human tendencies to predict neutral.

3.2 Performance on literary data across genres

Table 2 compares our model's predictions to the human gold-standard ratings for the *Fiction4* dataset's 4 genres.

Type	Model	Scalar	Overall	Hymns	Fairy tales	Prose	Poetry
Year				1798–1873	1837–1847	1952	1965
Human →	IRR ρ	✓	0.63	0.73	0.68	0.62	0.59
	IRR α	✓	0.67	0.72	0.68	0.61	0.58
↓ Dictionary	`vader`	✓	0.49	0.52	0.50	0.43	0.46
	`syuzhet`	✓	0.50	0.54	0.48	0.45	0.49
↓ Multiling.	`twitter-xlm`	×	0.55	0.50	0.52	0.57	0.58
	`xlm-roberta`	×	0.60	0.59	0.62	0.61	0.57
	`Sentiment Projection`	✓	**0.66**	**0.69**	0.66	**0.62**	**0.70**
↓ Danish	`danish-sentiment`	×	0.54	0.49	0.48	0.57	0.57
	`da-sentiment-base`	×	0.23	0.44	0.47	0.08	0.08
	`MeMo-BERT-SA`	×	0.47	0.63	**0.72**	0.26	0.16

Table 2: Spearman correlations in the *Fiction4* corpus *across genres*. From top to bottom: Publication years; then Inter Rater Reliability (human scores) per genre (Spearman's ρ and Krippendorff's α); then correlation between the human gold standard and models (Spearman's ρ). For VADER and Syuzhet scores, texts were automatically translated into English.

We evaluated all models on the full multilingual *Fiction4* corpus. For the dictionary-based tools (`VADER` and `Syuzhet`), originally Danish texts were translated into English (see subsubsection 2.2.1). Danish-specific models generally under-perform on genres that are (originally) in English (*Prose*, *Poetry*), which drags down their overall correlation scores. An outlier is `danish-sentiment`, which delivers relatively consistent results across both languages; however, it still falls short of `MeMo-BERT-SA` on the original Danish texts – most notably in the Fairy Tales genre.

Most Danish transformer-based models perform on par with (or worse than) dictionary-based models applied to English translations of the original Danish texts (e.g., Fairy tales & Hymns). Sentiment Projection, in contrast, achieves the highest correlation on every genre except Fairy tales – where `MeMo-BERT-SA` performed best, which aligns with its fine-tuning on Danish literary prose from H.C. Andersen's period. It performs especially well on Poetry, where other models struggle.

The genres that achieved the highest human IRR – like hymns, at IRR $\rho = 0.77$ – did not reflect in better results for most models. The second-best performing model, `xlm-roberta`, for example, placed second-to-last on hymns. Instead, Sentiment Projection meets or exceeds Inter Rater correlation (ρ) for all genres.

3.3 Performance on literary data across time and languages

Results for the multilingual performance assessment are presented in Table 3.

Type	**Model**	**Scalar**	**Multiling.** [Da + En]	**Danish set** [Da]	**English set** [En]	**Translated** [Da → En]
Human →	IRR ρ	✓	0.63	0.68	0.58	-
	IRR α	✓	0.67	0.71	0.60	-
↓ Dictionary	`vader`	✓	-	-	0.45	0.51
	`syuzhet`	✓	-	-	0.47	0.50
↓ Multiling.	`twitter-xlm`	×	0.55	0.50	0.58	0.56
	`xlm-roberta`	×	0.60	0.59	**0.60**	0.57
	`Sentiment Projection`	✓	**0.66**	**0.68**	**0.60**	**0.65***
↓ Danish	`danish-sentiment`	×	0.53	0.47	0.58	0.55
	`da-sentiment-base`	×	0.23	0.43	0.08	0.10
	`MeMo-BERT-SA`	×	0.48	0.67	0.25	0.24

Table 3: Spearman correlations in the *Fiction4* corpus *across languages*. Columns from left to right: Overall evaluation on **Multilingual** dataset (English and Danish); evaluation of the **Danish set** ($n = 2,800$); evaluation of the **English set** ($n = 3,500$); lastly, the evaluation of **Translated** set. On top, Inter Rater Reliability – Spearman's ρ and Krippendorff's α. The best model performance per setting is in bold, and the follow-up is underlined. ***** There might be minimal influx in correlation caused by the concept vector being defined by untranslated sentences that are included after translation.

Table 3 demonstrates that our Sentiment Projection model leads baselines in both multilingual and Danish-only evaluations. This gain likely reflects our use of a multilingual encoder for sentence embeddings and a "concept vector" defined over a multilingual corpus. Concretely, Sentiment Projection attains Spearman's $\rho = 0.68$ on the Danish subset (Fairytales + Hymns) versus $\rho = 0.58$ for the runner-up, and delivers a $\rho = 0.06$ absolute improvement in the overall multilingual setting.

We test our model for its generalization across time periods in Table 2, where danish hymns and fairytales represent historical language with texts from the 18-19th century. The Sentiment Projection model shows no signs of reduced performance when processing older texts and outperforms the follow-up model by $\rho = 0.12$ in the Hymns genre.

Notably, `twitter-xlm` model appears to perform slightly better on sentences translated to English than on their original Danish, as seen in Table 3. This may indicate that Google Translate renders language in updated, contemporaneous forms, similar to the Twitter data used for model training. We see the same tendency (surprisingly) for the `danish-sentiment` model, i.e., better

performance when Danish sentences were translated to English. In contrast, Sentiment Projection performs slightly better on the Danish set in its original form than when it is translated to English – which we consider validates its capacity to process older forms reliably.

3.4 Performance on literary and non-literary contemporary data

To make sure that our model does not overfit its sentiment vector to the in-context sentiment cues of the stories in the *Fiction4* corpus, we tested it against the *EmoBank* dataset – which indexes contemporary literary and non-literary data. All Multilingual and dictionary-based models were tested for their correlation with the human gold standard of the *EmoBank* dataset. The Sentiment Projection Model still achieved the highest overall correlation with human ratings. Although it shows a lower correlation for a few genres (i.a., Letters), it still appears to generalize well to contemporary out-of-training distribution data. It should be noted that the model outperforms the other models the most in the fiction genre, indicating that the sentiment vector may be slightly fine-tuned or overfit to fiction-specific sentiment indicators. While this can also be a drawback, it supports the idea that domain-specific sentiment analysis can be highly beneficial. For example, a sentiment analysis method for fiction should be sensitive to the specific sentiment cues (like omission, implicitness, concrete and object-based, etc.), rarer in other genres [3]. Feldkamp et al. [14] suggests that travel guides use similar mechanisms – sentiment is evoked through unsentimental, descriptive, and concrete detail. The fact that Sentiment Projection performs well also for both genres suggests it captures this kind of indirect sentiment expression.

	Scalar	Overall	Letters	Blog	Newspaper	Essays	Fiction	Travelguides
Human IRR α	✓	0.34	0.34	0.31	0.29	0.31	0.35	0.23
`vader`	✓	0.43	0.47	0.41	0.42	0.32	0.37	0.35
`syuzhet`	✓	0.46	0.47	0.37	0.42	0.37	0.43	0.37
`twitter-xlm`	×	0.64	**0.69**	**0.65**	0.61	**0.59**	0.57	0.48
`xlm-roberta`	×	0.65	0.68	**0.65**	0.65	0.58	0.56	0.49
`Sentiment Projection`	✓	**0.67**	0.62	0.61	**0.66**	0.53	**0.64**	**0.52**

Table 4: Spearman correlations on the `EmoBank` sentences ($n = 8,870$) *across domains*. On top: Inter Rater Reliability (Krippendorff's α).

4 Discussion & conclusions

As seen in Table 3 and 4, the proposed Sentiment Projection model performs on par with or better than the contemporary state-of-the-art methods. Moreover, Sentiment Projection allows for a *smooth continuous output*. In contrast, methods converting model output are not continuous in practice, but rather return noncontinuous tri-modal distributions (Figure 3). While both methods correlate highly with the human golden standard, approaching the inter-rater correlation, it appears that the *Sentiment Projection approach more closely resembles* the sentiment distribution of human ratings.

Furthermore, the Sentiment Projection method can be trained on multilingual data using a multilingual language model, allowing for a *language-agnostic sentiment prediction model* that also reliably handles historical variants. The Sentiment Projection was solely defined by its concept vector, based on sentences from the *Fiction4* dataset, half of which were in Danish, yet it still outperforms other models.

While this paper corroborates the findings of [13], showing that translation (even without a quality check) to English increases the similarity of human and transformer-model scores, it also

shows that this is not the case for Sentiment Projection, which performs slightly better on the original (Danish) sentences.

Finally, the workflow presented in Figure 2 has been used to design a sentiment model, but allows easy generalization to other concepts of choice. The method could also work for other emotional concepts, such as emotion recognition, language detection, or abstract concepts like a nature-to-industry gradient. We encourage curious readers to search for inspiration for potential vectors in *Linear Representation Hypothesis* [25] and *Steering Vector* [35] literature. Due to the flexible nature of the algorithm, there is no rigid lower boundary on the number of training points required for a stable vector, although the chances of over-representing the non-concept context of training sentences naturally increase as the number of sentences decreases. A future empirical investigation of the stability of the vector when using smaller training sets would be useful.

Acknowledgements

References

[1] Allaith, Ali, Degn, Kirstine, Conroy, Alexander, Pedersen, Bolette, Bjerring-Hansen, Jens, and Hershcovich, Daniel. "Sentiment Classification of Historical Danish and Norwegian Literary Texts". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, ed. by Tanel Alumäe and Mark Fishel. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 324–334. URL: https://aclanthology.org/2023.nodalida-1.34 (visited on 05/07/2024).

[2] Batanović, Vuk, Cvetanović, Miloš, and Nikolić, Boško. "A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts". In: *PLoS ONE 15*, no. 11 (Nov. 2020). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0242050. (Visited on 05/07/2024).

[3] Bizzoni, Yuri and Feldkamp, Pascale. "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*, ed. by Valentina Pyatkin, Daniel Fried, Elias Stengel-Eskin, Alisa Liu, and Sandro Pezzelle. Malta: Association for Computational Linguistics, Mar. 2024, pp. 54–61. URL: https://aclanthology.org/2024.unimplicit-1.5.

[4] Bizzoni, Yuri and Feldkamp, Pascale. "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*. Tokyo, Japan: Association for Computational Linguistics, 2023, pp. 219–226. URL: https://rootroo.com/downloads/nlp4dh_iwclul_proceedings.pdf.

[5] Bizzoni, Yuri, Moreira, Pascale, Thomsen, Mads Rosendahl, and Nielbo, Kristoffer. "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*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 11–18. URL: https://aclanthology.org/2023.wassa-1.2.

[6] Bizzoni, Yuri, Peura, Telma, Nielbo, Kristoffer, and Thomsen, Mads. "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*. Taipei, Taiwan: Association for Computational Linguistics, Nov. 2022, pp. 31–41. URL: https://aclanthology.org/2022.nlp4dh-1.5.

[7] Bollen, Johan, Mao, Huina, and Zeng, Xiaojun. “Twitter mood predicts the stock market”. In: *Journal of Computational Science 2*, no. 1 (Mar. 2011), pp. 1–8. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2010.12.007. (Visited on 05/07/2024).

[8] Booth, Wayne C. *The Rhetoric of Fiction*. English. 2nd edition. Chicago: University of Chicago Press, Feb. 1983. ISBN: 978-0-226-06558-8.

[9] Buechel, Sven and Hahn, Udo. “EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 578–585. URL: https://aclanthology.org/E17-2092/.

[10] De Bruyne, Luna, Singh, Pranaydeep, De Clercq, Orphee, Lefever, Els, and Hoste, Veronique. “How Language-Dependent is Emotion Detection? Evidence from Multilingual BERT”. In: *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, ed. by Duygu Ataman, Hila Gonen, Sebastian Ruder, Orhan Firat, Gözde Gül Sahin, and Jamshidbek Mirzakhalov. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 76–85. DOI: 10.18653/v1/2022.mrl-1.7.

[11] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. “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)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[12] Enevoldsen, Kenneth. “Asent: Fast, flexible and transparent sentiment analysis”. Version 0.4.0. 2022. URL: https://github.com/KennethEnevoldsen/asent.

[13] Feldkamp, Pascale, Kostkan, Jan, Overgaard, Ea, Jacobsen, Mia, and Bizzoni, Yuri. “Comparing Tools for Sentiment Analysis of Danish Literature from Hymns to Fairy Tales: Low-Resource Language and Domain Challenges”. In: *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, ed. by Orphée De Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 186–199. URL: https://aclanthology.org/2024.wassa-1.15 (visited on 09/20/2024).

[14] Feldkamp, Pascale, Lindhardt, Ea Overgaard, Nielbo, Kristoffer L., and Bizzoni, Yuri. “Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. 2024, pp. 681–706.

[15] Hutto, C. and Gilbert, Eric. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media 8*, no. 1 (May 2014). Number: 1, pp. 216–225. ISSN: 2334-0770. DOI: 10.1609/icwsm.v8i1.14550. (Visited on 07/10/2025).

[16] Jockers, Matthew. “A Novel Method for Detecting Plot”. en-US. Matthew L. Jockers Blog. 2014. URL: https://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/ (visited on 09/12/2022).

[17] Jockers, Matthew L. “Syuzhet: Extract Sentiment and Plot Arcs from Text”. 2015. URL: https://github.com/mjockers/syuzhet.

[18] Kim, Been, Wattenberg, Martin, Gilmer, Justin, Cai, Carrie, Wexler, James, Viegas, Fernanda, et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)”. In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[19] Al-Laith, Ali, Conroy, Alexander, Bjerring-Hansen, Jens, and Hershcovich, Daniel. “Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, May 2024, pp. 4811–4819. URL: https://aclanthology.org/2024.lrec-main.431/ (visited on 04/30/2025).

[20] Al-Laith, Ali, Degn, Kirstine, Conroy, Alexander, Pedersen, Bolette, Bjerring-Hansen, Jens, and Hershcovich, Daniel. “Sentiment Classification of Historical Danish and Norwegian Literary Texts”. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, ed. by Tanel Alumäe and Mark Fishel. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 324–334. URL: https://aclanthology.org/2023.nodalida-1.34/.

[21] Lauridsen, Gustav Aarup, Dalsgaard, Jacob Aarup, and Svendsen, Lars Kjartan Bacher. “SENTIDA: A New Tool for Sentiment Analysis in Danish”. en. In: *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift 4*, no. 1 (Sept. 2019). Number: 1, pp. 38–53. ISSN: 2446-0591. URL: https://tidsskrift.dk/lwo/article/view/115711 (visited on 05/08/2024).

[22] Li, Belinda Z., Nye, Maxwell, and Andreas, Jacob. “Implicit Representations of Meaning in Neural Language Models”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 1813–1827. DOI: 10.18653/v1/2021.acl-long.143.

[23] Lukes, Jan and Søgaard, Anders. “Sentiment analysis under temporal shift”. In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, ed. by Alexandra Balahur, Saif M. Mohammad, Veronique Hoste, and Roman Klinger. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 65–71. DOI: 10.18653/v1/W18-6210. (Visited on 07/16/2025).

[24] Park, Kiho, Choe, Yo Joong, and Veitch, Victor. “The Linear Representation Hypothesis and the Geometry of Large Language Models”. arXiv:2311.03658 [cs]. July 2024. DOI: 10.48550/arXiv.2311.03658. URL: http://arxiv.org/abs/2311.03658 (visited on 04/16/2025).

[25] Peterson, Joshua C., Chen, Dawn, and Griffiths, Thomas L. “Parallelograms revisited: Exploring the limitations of vector space models for simple analogies”. In: *Cognition 205* (2020), p. 104440. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2020.104440.

[26] Reagan, Andrew J, Mitchell, Lewis, Kiley, Dilan, Danforth, Christopher M, and Dodds, Peter Sheridan. “The Emotional Arcs of Stories Are Dominated by Six Basic Shapes”. en. In: *EPJ Data Science 5*, no. 1 (Dec. 2016), pp. 1–12. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-016-0093-1. (Visited on 09/06/2022).

[27] Rebora, Simone, Lehmann, Marina, Heumann, Anne, Ding, Wei, and Lauer, Gerhard. "Comparing ChatGPT to Human Raters and Sentiment Analysis Tools for German Children's Literature". In: *Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023*, ed. by Artjoms Sela, Fotis Jannidis, and Iza Romanowska. Vol. 3558. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 333–343. URL: https://ceur-ws.org/Vol-3558/paper3340.pdf.

[28] Reimers, Nils and Gurevych, Iryna. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: http://arxiv.org/abs/1908.10084.

[29] Schmidt, Thomas and Burghardt, Manuel. "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*, ed. by Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Santa Fe, New Mexico: Association for Computational Linguistics, Aug. 2018, pp. 139–149. URL: https://aclanthology.org/W18-4516 (visited on 05/08/2024).

[30] Schmidt, Thomas, Dennerlein, Katrin, and Wolff, Christian. "Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays". en. In: *Fabrikation von Erkenntnis: Experimente in den Digital Humanities* - (2021). DOI: 10.26298/MELUSINA.8F8W-Y749-UDLF. (Visited on 05/07/2024).

[31] Swafford, Annie. "Problems with the Syuzhet Package". en. In: *Anglophile in Academia: Annie Swafford's Blog* (Mar. 2015). URL: https://annieswafford.wordpress.com/2015/03/02/syuzhet/ (visited on 09/08/2022).

[32] Tsao, Hsiu-Yuan, Chen, Ming-Yi, Lin, Hao-Chiang Koong, and Ma, Yu-Chun. "The asymmetric effect of review valence on numerical rating: A viewpoint from a sentiment analysis of users of TripAdvisor". In: *Online Information Review 43*, no. 2 (Jan. 2018). Publisher: Emerald Publishing Limited, pp. 283–300. ISSN: 1468-4527. DOI: 10.1108/OIR-11-2017-0307. (Visited on 05/07/2024).

[33] Vishnubhotla, Krishnapriya, Hammond, Adam, Hirst, Graeme, and Mohammad, Saif. "The Emotion Dynamics of Literary Novels". In: *Findings of the Association for Computational Linguistics: ACL 2024*, ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2557–2574. DOI: 10.18653/v1/2024.findings-acl.150.

[34] Vu, Thuy and Parker, D. Stott. "K-Embeddings: Learning Conceptual Embeddings for Words using Context". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1262–1267. DOI: 10.18653/v1/N16-1151.

[35] Wehner, Jan, Abdelnabi, Sahar, Tan, Daniel, Krueger, David, and Fritz, Mario. "Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models". arXiv:2502.19649 [cs]. Mar. 2025. DOI: 10.48550/arXiv.2502.19649. URL: http://arxiv.org/abs/2502.19649 (visited on 06/26/2025).

[36] Wilson, Theresa, Wiebe, Janyce, and Hoffmann, Paul. “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis”. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, ed. by Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 347–354. URL: https://aclanthology.org/H05-1044 (visited on 05/07/2024).

[37] Xu, Yuemei, Cao, Han, Du, Wanze, and Wang, Wenqing. “A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations”. en. In: *Data Science and Engineering 7*, no. 3 (Sept. 2022), pp. 279–299. ISSN: 2364-1185, 2364-1541. DOI: 10.1007/s41019-022-00187-3. (Visited on 05/06/2024).

[38] Zehe, Albin, Becker, Martin, Hettinger, Lena, Hotho, Andreas, Reger, Isabella, and Jannidis, Fotis. “Prediction of Happy Endings in German Novels Based on Sentiment Information”. en. In: *Interactions between Data Mining and Natural Language Processing*, ed. by Peggy Cellier, Thierry Charnois, Andreas Hotho, Stan Matwin, Marie-Francine Moens, and Yannick Toussaint. Riva del Garda, 2016, pp. 9–16.

[39] Zhao, Haiyan, Zhao, Heng, Shen, Bo, Payani, Ali, Yang, Fan, and Du, Mengnan. “Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution”. In: *arXiv preprint arXiv:2410.00153* (2024).

A Models

Type		Shorthand, Modelname & URLs
↓ Encoder		
	Shorthand	Sentiment Projection
	Name	`Sentiment Projection using paraphrase-multilingual-mpnet-base-v2`
	URL	https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
↓ Multiling.		
	Shorthand	twitter-xlm
	Name	`cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual`
	URL	https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual
	Shorthand	xlm-roberta
	Name	`cardiffnlp/xlm-roberta-base-sentiment-multilingual`
	URL	https://huggingface.co/cardiffnlp/xlm-roberta-base-sentiment-multilingual
↓ Danish		
	Shorthand	danish-sentiment
	Name	`vesteinn/danish_sentiment`
	URL	https://huggingface.co/vesteinn/danish_sentiment
	Shorthand	da-sentiment-base
	Name	`alexandrainst/da-sentiment-base`
	URL	https://huggingface.co/alexandrainst/da-sentiment-base
	Shorthand	MeMo-BERT-SA
	Name	`MiMe-MeMo/MeMo-BERT-SA`
	URL	https://huggingface.co/MiMe-MeMo/MeMo-BERT-SA

Table 5: Full model names & details.

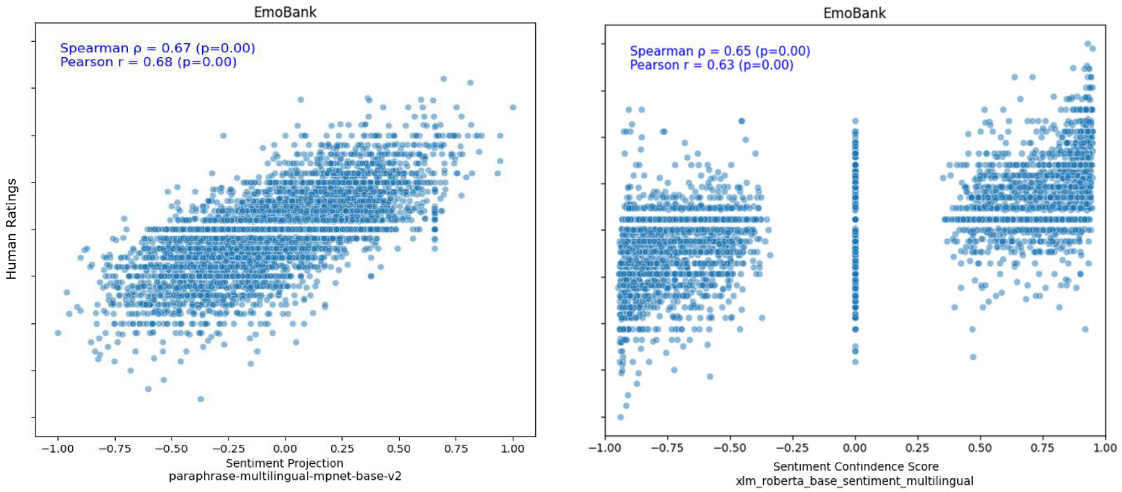

Figure 4: Scatterplot of Sentiment Projection `xlm-roberta` for *EmoBank* Data.

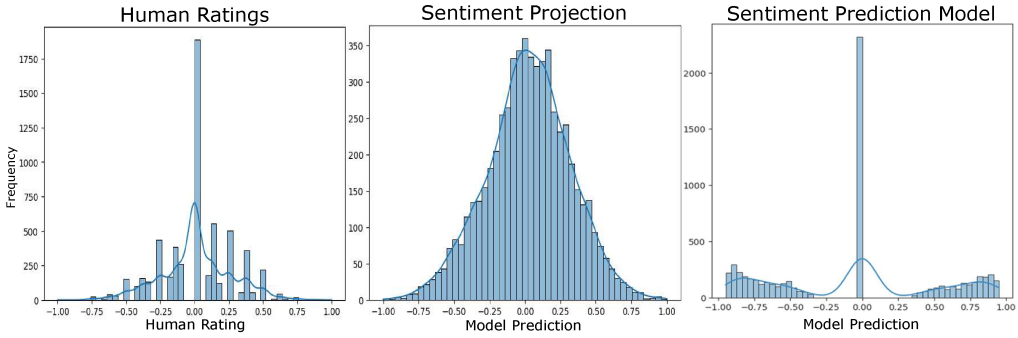

Figure 5: Histograms of respectively Human raters, sentiment projection model and `xlm-roberta`'s predictions for the *Fiction4* test-set. This plot should be interpreted in conjunction with Figure 3 and Figure 4. It visualizes that the `xlm-roberta` model follows the human trend of predicting completely neutral sentences. The Sentiment Projection predicts mostly neutral sentences, as hoped, but follows a bell-curve that becomes visible in human ratings, as the number of raters increases, see Figure 6.

B Score distribution

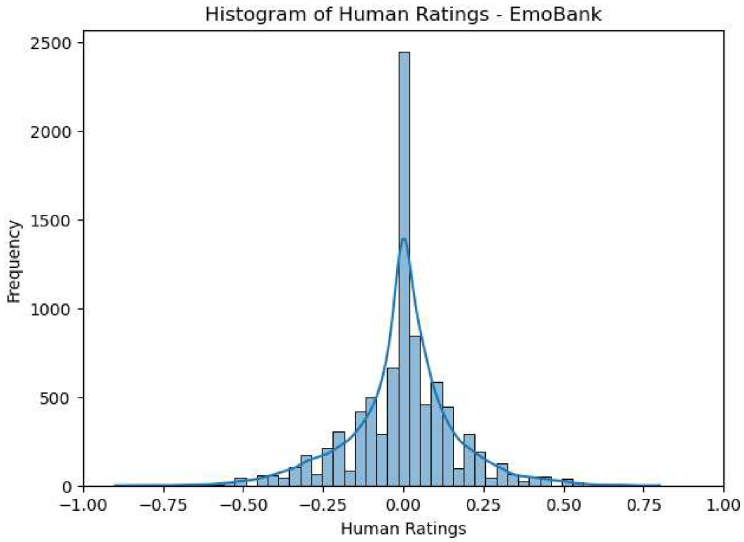

Figure 6: Histogram of human ratings. As ratings become the average of 10 raters, it approaches a more continuous bell-shaped form, in comparison to the 3-rater average depicted in the *Human Rating* plot in Figure 5.

C Algorithm

The following algorithm formally describes the procedure for defining and applying a concept vector by using labeled sentence embeddings.

Algorithm 1 Concept Vector Projection

Input:
$\mathcal{M}$ = Language Model
$\mathcal{S}$ = A set of categorically labeled sentences $s_i \in \{\text{positive}^+, \text{negative}^-, \text{neutral}^\varnothing, \text{unknown}^?\}$
Output:
$\hat{\mathbf{v}}$ = Concept vector
score(s_i) = projection scores for unknown sentences
Computation:

Embed all sentences: $\mathbf{e}_i = \mathcal{M}(s_i)$
$P^+ \leftarrow \{\mathbf{e}_i \mid s_i = \text{positive}\}$
$N^- \leftarrow \{\mathbf{e}_i \mid s_i = \text{negative}\}$
Compute means: $\vec{\mu_{s^+}} = \text{mean}(P^+)$, $\vec{\mu_{s^-}} = \text{mean}(N^-)$
Compute concept vector: $\vec{\mathbf{v}} = \vec{\mu_{s^+}} - \vec{\mu_{s^-}}$
Normalize: $\hat{\mathbf{v}} = \frac{\vec{\mathbf{v}}}{\|\vec{\mathbf{v}}\|}$
for each embedding $\mathbf{e}_i$ **do**
    score$(s_i) = \mathbf{e}_i \cdot \hat{\mathbf{v}}$ // Embedding projection
end for