Is Sentiment Banana-Shaped? Exploring the Geometry and Portability of Sentiment Concept Vectors

Laurits Lyngbaek*, **Pascale Feldkamp***, **Yuri Bizzoni**,
Kristoffer L. Nielbo, **Kenneth Enevoldsen**.
Aarhus University, Aarhus, Denmark
*Shared First Authorship

Abstract

Use cases of sentiment analysis in the humanities often require contextualized, continuous scores. Concept Vector Projections (CVP) offer a recent solution: by modeling sentiment as a direction in embedding space, they produce continuous, multilingual scores that align closely with human judgments. Yet the method's portability across domains and underlying assumptions remain underexplored. We evaluate CVP across genres, historical periods, languages, and affective dimensions, finding that concept vectors trained on one corpus transfer well to others with minimal performance loss. To understand the patterns of generalization, we further examine the linearity assumption underlying CVP. Our findings suggest that while CVP is a portable approach that effectively captures generalizable patterns, its linearity assumption is approximate, pointing to potential for further development. Code available at: github.com/lauritswl/representation-transfer

1 Introduction and Related Works

Sentiment Analysis approaches to data in the Humanities often need continuous sentiment scores to develop meaningful models of texts, for tasks such as tracing the "sentiment arc" of a story (Jockers, 2014; Reagan et al., 2016; Zehe et al., 2016; Bizzoni et al., 2023), gauging sentiment fluctuations in news (Daudert, 2021) or modeling changes in online discourse (Xie and He, 2025), but existing tools struggle to capture the necessary nuances effectively. Many dictionary-based methods are continuous, but struggle with extended context, whereas Transformer models produce binary or trinary outputs that only approximate continuous sentiments through post-hoc adjustments (Bizzoni and Feldkamp, 2023; Lyngbaek et al., 2025).

A recent alternative (Lyngbaek et al., 2025) uses a projection-based method in a homogeneous semantic space to generate continuous sentiment scores that align with human judgments and match or surpass Transformer-based methods on literary data, while producing smoother distributions. This approach, called Concept Vector Projection (CVP), rests on the "linear representation hypothesis" (Park et al., 2024): the idea that semantic concepts, such as sentiments, can be represented linearly in embedding space (Wehner et al., 2025; Vu and Parker, 2016; Li et al., 2021; Zhao et al., 2024). Under this idea, a given semantic concept corresponds to a *direction* in the embedding space, so that moving further along this direction increases its intensity (see Figure 1).

While studies have validated this idea at various levels of abstraction (Lyngbaek et al., 2025; Wehner et al., 2025; Vu and Parker, 2016; Li et al., 2021; Zhao et al., 2024), its *portability* across different data domains and semantic dimensions remains underexplored. Literary texts, blogs, newspapers, and social media differ in style and affective expression (Feldkamp et al., 2024; Vishnubhotla et al., 2024),[1] and language or period differences can complicate the transfer. The general trend in sentiment analysis has been to assume non-portability and train or fine-tune specialized models for specific domains, languages and historical variants (Allaith et al., 2023; Schmidt and Burghardt, 2018) – models that are then difficult to adapt for other use cases.

In this work, we test the CVP's portability across three datasets, spanning genres (social media to letters), periods (1798-2013), and languages (English and Danish). After presenting the Data (section 2) and Methods (section 3), we test the CVP through several cross-dataset experiments (subsection 4.1) to assess whether the resulting scores retain their alignment with human judgments. We also explore

[1] How much domains differ varies. For example, if 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 (Feldkamp et al., 2024).

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 146–160
March 29, 2026 ©2026 Association for Computational Linguistics

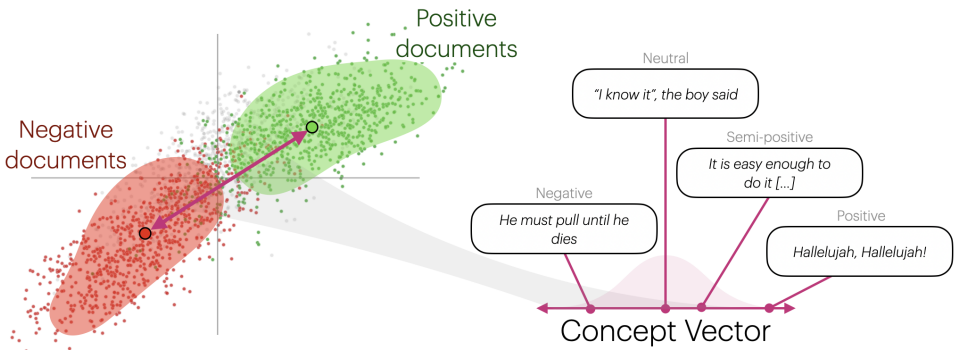

Figure 1: A visualization of how the Concept Vector Projection is constructed. It shows how to construct a positive-negative concept vector to predict sentiment in an unlabeled corpus in a continuous space.

the portability of the CVP beyond valence – to related affective dimensions, such as arousal and dominance (subsection 4.2), and consider whether imperfect linearity in the projections might be the cause of some of the method's inaccuracies (subsection 4.3).

2 Datasets

To represent diversity across the literary and non-literary domains, we select three datasets that span different genres, domains, languages, and historical periods, all using a human-annotated continuous scale.

Emobank (Buechel and Hahn, 2017) contains sentences from the MASC dataset annotated according to the Valence-Arousal-Dominance (VAD) scheme (Mehrabian and Russell, 1974). The dataset includes: Letters, Blog, Newspaper, Essays, Fiction, and Travel guides.

Facebook (Preoţiuc-Pietro et al., 2016) consists of status updates collected by (Kosinski et al., 2013) and annotated for valence and arousal.

Fiction4 (Feldkamp et al., 2024) comprises literary texts spanning four genres and two languages (English/Danish) from the 19th and 20th centuries. It consists of three main authors – Sylvia Plath (poetry), Ernest Hemingway (prose), and H.C. Andersen (fairytales) – and hymns from Danish official church hymnbooks (published 1798-1873). Two or more human annotators scored each sentence (/line, for poetry) for valence.[2]

[2]Although lower than Facebook posts, IRR for Fiction4 (α=0.67) is high for continuous annotations of literary texts. Humans rarely reach $\alpha > 0.80$ for polarity tagging on *non-literary* texts (Wilson et al., 2005) and achieve lower IRR for continuous scales on literary texts (Batanović et al., 2020; Rebora et al., 2023).

Dataset	Period	Sentences	Kripp. α (Scale)
EmoBank	1990–2008	10,062	
Valence			.34 (1-5)
Arousal			.25 (1-5)
Dominance			.22 (1-5)
Facebook	2012–2013	2,895	
Valence			.72 (1-9)
Arousal			.82 (1-9)
Fiction4	1798–1965	6,300	
Valence			.67 (0-10)

Table 1: Summary of annotated corpora. We report sentence counts, average length, and inter-rater agreement (α). The total number of sentences considered is $n = 19,257$. Full breakdown of subgenres (in Fiction4 and Emobank) and number of annotators in Appendix D.

3 Methods

To construct the concept vector, we follow the approach introduced by (Lyngbaek et al., 2025), where a pre-trained sentence-embedding model $\mathbf{M}$ embeds a set of source$_{negative}$ and target$_{positive}$ exemplar sentences. We compute the mean embeddings of source$_{negative}$ and target$_{positive}$ examples and define the concept vector $\hat{\mathbf{v}}$ as the unit vector of the difference between mean embeddings. The assumption is that this averaging will reduce non-sentiment information to Gaussian noise with a mean of zero, leaving the sentimental signal behind (Kim et al., 2018; Zhao et al., 2024). With this method, we score a sentence s by projecting its embedding onto the concept vector $\hat{\mathbf{v}}$ via the dot product $\mathbf{M}(s) \cdot \hat{\mathbf{v}}$, yielding a continuous sentiment score. We normalize the scores using a z-score normalization. We define the details for the CVP algorithm in Appendix B.[3] To define source$_{negative}$ and target$_{positive}$, we set sentiment thresholds relative to each corpus' valence distribution. Sentences

[3]Implementation available at representation-transfer

at least one standard deviation above the mean are positive, sentences below by one standard deviation are negative, and the rest are neutral (for the formalization, see Appendix A). We estimate *concept vectors* from these positive–negative contrasts, capturing the extremes rather than absolute ratings. This approach yields comparable sentiment contrasts across datasets with different scales and distributions. For testing linearity, we created three concept vectors: positive–negative, negative–neutral, and neutral–positive.

3.1 Model

To allow for comparability with previous works (Lyngbaek et al., 2025), we use the embedding model paraphrase-multilingual-mpnet[4] (Reimers and Gurevych, 2019), a 278M-parameter model based on a mean-pooled BERT architecture optimized for sentence similarity via Siamese and Triplet networks. This model is notable for its multilingual capabilities, previous performance (Lyngbaek et al., 2025), and excellent size-to-performance ratio.[5]

4 Results

4.1 Portability

Our results show that the projection method is robust: continuous valence scores remain well-aligned to human scores across all three datasets and their constituent subgenres (Table 2), suggesting that the approach captures generalizable sentiment patterns beyond the idiosyncrasies of literary, journalistic, or social media language. It highlights the portability of continuous sentiment scoring across genres, which can be crucial for research spanning multiple text types or for investigating historical and contemporary corpora side by side.

Dataset	Correlation, when trained on:		
	Fiction4	**Emobank**	**Facebook**
Fiction4	**0.66**	0.65	0.64
Emobank	0.67	**0.70**	0.66
Facebook	0.66	0.66	**0.68**

Table 2: Spearman correlations between human and projected valence scores across corpora. Values indicate correlations when trained on the indicated corpus (columns) and tested on itself or another corpus (rows).

[4]sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[5]A larger model may increase model correlation with human scores at the expense of computation budget.

	Emobank	**Facebook**	**Fiction4**
Valence	.71±.02(.70)	.70±.02(.68)	.66±.02(.66)
Arousal	.36±.02(.42)	.65±.02(.67)	
Dominance	.35±.01(.37)		

Table 3: Cross-validation of Spearman correlations between CVP scores and human scores for valence, arousal, and dominance per corpus. The scores are the mean correlation obtained from a five fold analysis, with a standard deviation notated by ±. The score parenthesis indicates the Spearman correlation obtained when no split was conducted. Only Emobank has human scores of all V-A-D labels.

4.2 Beyond Valence

To test the CVP's ability to generalize beyond valence – which refers to the positivity/negativity spectrum – we tested the approach on semantic properties associated with valence in sentiment analysis: arousal and dominance. Arousal refers to the intensity of the concept conveyed by a given word (*ecstatic* and *serene* are both positive, but the first word elicits a higher arousal); dominance refers to the amount of control associated with a term (*angry* and *helpless* are both negative, but the first word has more dominance). We find that CVP generalizes well for these subtler concepts (Table 3) with similarly continuous distributions (see Appendix F), without reaching the performance achieved on valence.

4.3 Linearity assumption

CVP treats sentiment as linear in embedding space: negative and positive extremes form the main axis, with neutral sentences in the middle. We create similar vectors with negative–neutral and neutral–positive extremes, and examine the cosine similarity between all three vectors (Figure 3). Across corpora, the negative–positive axis aligns most strongly with the other two axes, consistent with a geometrical interpretation that neutral texts are located close to the axis, but vary along an undefined semantic dimension, so that the three vectors will form a triangle outlining the centroids of the three classes. We explore the geometry of Fiction4's valence space by creating a two-dimensional basis that visualizes the data. We define our first dimension as the negative-positive vector, $\mathbf{v}_{np}$. The second semantic dimension we define as the neutral-component, $\mathbf{v}_{nc}$. The neutral-component vector captures the remaining semantic information encoded in the neutral centroid.

To define the neutral component, we treat the

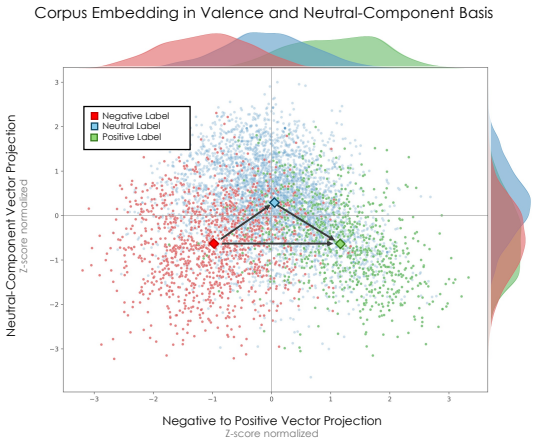

Figure 2: Scatterplot visualizing the Fiction4 embeddings projected onto the Fiction4 Pos-Neg Sentiment vector and the corresponding Neutral-Component. The Margin plots are Kernel Density Estimations of the label distributions. All dimensions are Z-score normalized to make the projections interpretable.

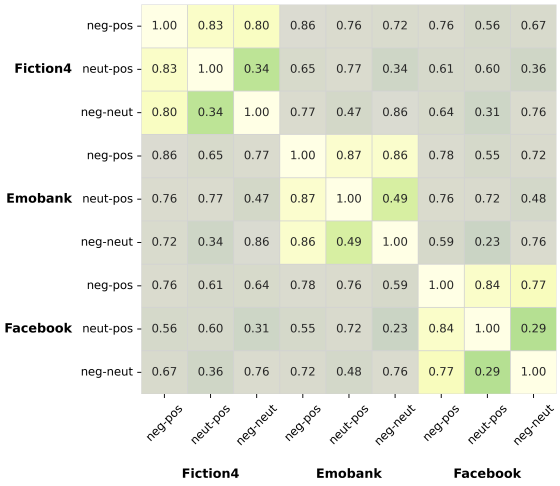

Figure 3: Cosine similarity between Concept Vectors for each corpus (values in each cell). Internal correlations among neg-pos, neut-pos, and neg-neut pairs are strong, with neut-pos and neg-neut closer to neg-pos, reflecting a centrality of the negative–positive axis across corpora.

problem geometrically in an affine space. Because an affine space has no natural origin, projections cannot be applied directly to centroids. Instead, we work with *difference vectors*, which encode relative positions between centroids. Let $\mathbf{v}_{np}$ denote the vector from the negative to the positive centroid and define the corresponding unit direction

$$\hat{\mathbf{v}}_{np} = \frac{\mathbf{v}_{np}}{\|\mathbf{v}_{np}\|}.$$

We also define the vector from the negative to the neutral centroid as $\mathbf{v}_{nn}$. Projecting $\mathbf{v}_{nn}$ onto $\hat{\mathbf{v}}_{np}$ gives the scalar projection

$$k = \mathbf{v}_{nn} \cdot \hat{\mathbf{v}}_{np}.$$

This scalar specifies the relative position of the neutral centroid along the negative–positive axis, yielding the projected component

$$k\,\hat{\mathbf{v}}_{np}.$$

Finally, we construct the *neutral component vector* by removing this projected information:

$$\mathbf{v}_{nc} = \mathbf{v}_{nn} - k\,\hat{\mathbf{v}}_{np}.$$

This residual vector represents the component of the neutral direction that is orthogonal to the sentiment axis. We use this constructed basis to visualize the geometric structure of sentiment embeddings, as seen in Figure 2. This result aligns with the high-dimensional cosine-similarity observed in Figure 3.

We see that our embeddings tend to be linear, but that neutral embeddings encode spurious information that remains unaccounted for in the sentiment direction. This property gives the centroids a triangular shape – and structures the Fiction4 embeddings as a banana-shaped manifold.

5 Discussion & conclusions

We find that Concept Vector Projections transfer well across genres, periods, and languages — a vector derived from a corpus including 19th-century Danish hymns predicts sentiment in contemporary Facebook posts nearly as well as one trained on in-domain data. This portability suggests CVP captures generalizable properties of how sentiment is encoded in embedding space, rather than domain-specific patterns. The approach also extends to arousal and dominance, though with reduced performance, consistent with valence being the most reliable and consistent of the three affective dimensions (Warriner et al., 2013). For researchers working with historical or low-resource corpora, this means domain-specific training data may not be necessary – as suggested in Lyngbaek et al. (2025).

Our geometric analysis shows that the linearity assumption is approximate: neutral sentences do not lie exactly on the positive–negative axis but form a continuous, banana-shaped curve. This suggests neutrality encodes semantic content beyond the absence of valence – a property that future methods might exploit.

Limitations

While the goal of this work, is not to explore how to optimize the performance of the CVP, but examine its implications, we only examine one model to ensure comparability with Lyngbaek et al. (2025). Further analysis should explore alternative models as indicate evidence suggests that newer models like EmbeddingGemma (Vera et al., 2025) might surpass the one currently used.

To examine the CVP ability to generalize to the related concepts arousal and dominance, we utilize the dataset itself as the source dataset for deriving the concept vector. This likely leads to a modest overestimation of the correlation as we see for valence in Table 2 and arousal in Appendix G.

Our cross-lingual evaluation, while leveraging a multilingual embedding model, is restricted to Danish and English. These languages, though differing in resource availability, belong to the same language family; generalization to typologically distinct languages remains untested.

Acknowledgments

This work was partially supported by the Danish National Research Foundation (DNRF193) through TEXT: Center for Contemporary Cultures of Text, Aarhus University.

References

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.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

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, 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.

Sven Buechel and Udo Hahn. 2017. 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*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Tobias Daudert. 2021. Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, 230:107389.

Pascale Feldkamp, Ea Overgaard Lindhardt, Kristoffer L. Nielbo, and Yuri Bizzoni. 2024. Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond. In *Proceedings of the Computational Humanities Research Conference*, volume 3834 of *CEUR Workshop Proceedings*, pages 681–706.

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

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018.

Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. 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)*, pages 1813–1827, Online. Association for Computational Linguistics.

Laurits Lyngbaek, Pascale Feldkamp, Yuri Bizzoni, Kristoffer L. Nielbo, and Kenneth Enevoldsen. 2025. Continuous Sentiment Scores for Literary and Multilingual Contexts. In *Anthology of Computers and the Humanities*, volume 3, pages 480–497, Luxembourg. ACH.

Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology.* the MIT Press.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. *arXiv preprint*. ArXiv:2311.03658 [cs].

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

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, Marina Lehmann, Anne Heumann, Wei Ding, and Gerhard Lauer. 2023. 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*, volume 3558 of *CEUR Workshop Proceedings*, pages 333–343. CEUR-WS.org.

Nils Reimers and Iryna Gurevych. 2019. 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.

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.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

Krishnapriya Vishnubhotla, Adam Hammond, Graeme Hirst, and Saif Mohammad. 2024. The emotion dynamics of literary novels. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2557–2574, Bangkok, Thailand. Association for Computational Linguistics.

Thuy Vu and D. Stott Parker. 2016. 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*, pages 1262–1267, San Diego, California. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models. *arXiv preprint*. ArXiv:2502.19649 [cs].

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.

Haihua Xie and Miao He. 2025. Tracking Fine-Grained Public Opinions: Two Datasets from Online Discourse on Trending Topics. *Mathematics*, 13(21):3433.

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.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. 2024. Beyond single concept

vector: Modeling concept subspace in llms with gaussian distribution. *arXiv preprint arXiv:2410.00153*.

A Vector polarity selection procedure

For each text unit i in a corpus, we compute its valence score v_i. Let μ and σ denote the mean and standard deviation of valence scores across the corpus. Polarity labels are assigned as follows:

$$\text{label}_i = \begin{cases} \text{positive}, & \text{if } v_i \geq \mu + \sigma \\ \text{negative}, & \text{if } v_i \leq \mu - \sigma \\ \text{neutral}, & \text{otherwise} \end{cases}$$

This scheme assigns a label based on deviation from the corpus mean by one standard deviation.

B CVP 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:
M = Language Model
S = A set of sentences s_i, labeled via mean ± SD thresholds for valence: $s_i \in \{\text{positive}^+, \text{negative}^-, \text{neutral}^\varnothing, \text{unknown}^?\}$
Output:
$\hat{\mathbf{v}}$ = Concept vector
$\text{score}(s_i)$ = projection scores for unknown sentences
Computation:

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

C Performance baseline

To contextualize the correlations between the CVP and human scores, we also include the correlations between a transformer-based model and human scores. We choose the best-performing model in Lyngbaek et al. (2025), the multilingual cardiffnlp/xlm-roberta-base-sentiment-multilingual (here abbreviated xlm-R-b)[6], which is an xlm-roberta model finetuned for sentiment on Twitter data (Barbieri et al., 2022). The model's output was transformed using its confidence scores, consistent with the approach in Lyngbaek et al. (2025) and Bizzoni and Feldkamp (2023). We do not compare to continuous dictionary-based sentiments approaches like VADER, but for a comparison against these methods, we refer to Lyngbaek et al. (2025).

Note that while the **xlm-R-b** model performs better than the Concept Vector Projection on Facebook data in terms of Spearman's ρ (see Table 2), the distributions of these scores remain pseudo-trinary (see the Figure 4), unlike the distribution of the Concept Vector Projection's scores (Figure 5).

[6] https://huggingface.co/cardiffnlp/xlm-roberta-base-sentiment-multilingual

Dataset	Subcategory	xlm-R-b	Correlation, when trained on:		
			Fiction4	Emobank	Facebook
Fiction4	*overall*	0.60	**0.66**	0.65	0.64
	fairytales	0.62	**0.67**	0.64	0.61
	hymns	0.59	**0.67**	0.66	0.62
	poetry	0.57	**0.72**	0.71	0.68
	prose	0.61	**0.64**	0.62	0.63
Emob.	*overall*	0.65	0.67	**0.70**	0.66
	SemEval	0.64	0.66	**0.71**	0.65
	blog	0.65	0.64	0.68	**0.69**
	essays	0.58	0.59	**0.63**	0.55
	fiction	0.56	0.67	**0.69**	0.67
	letters	0.68	0.68	**0.71**	0.66
	newspaper	0.65	0.67	**0.69**	0.65
	travel-guides	0.49	0.56	0.58	**0.59**
FB	*overall*	0.74	0.66	0.66	**0.68**

Table 4: Correlations with human and projected valence scores across corpora. Values indicate correlations when trained on the indicated corpus (columns) and tested on the datasets overall and across subgenres (rows). Correlation of the transformer-based model and human score is indicated in column **xlm-R-b**.

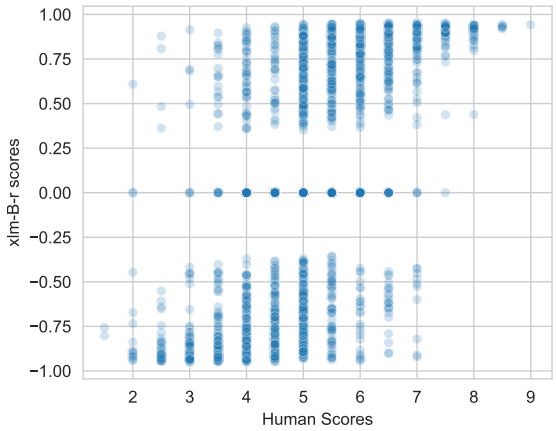

Figure 4: Correlation (Spearman's ρ) between transformer model (xlm-R-b) and human scores in the Facebook dataset.

D Data details

A note on *EmoBank* categories: *Essays* include, i.e., "A Brief History of Steel in Northeastern Ohio". *Fiction* comprises prose pieces, i.e., Richard Harding's "A Wasted Day" or the SciFi story "Captured Moments". *Newspapers* contain reports and longer reportages. *Travel Guides* include both local histories and reflective pieces (e.g., "Dublin and the Dubliners").[7]

[7]See the full MASC corpus at: https://anc.org/data/masc/corpus/browse-masc-data/

Dataset	Period	N annotations	N words	$\bar{x}$ words/sentence	N annotators	Krippendorff's α		
						V	A	D
→ *Facebook*	2012-2013	2,895	46,868	16.19	2	.72	.82	-
↓ *EmoBank*	1990-2008	10,062	151,259	15.03	10	.34	.25	.22
Letters		1,413	21,639	15.31	10	.35	.25	.25
Blog		1,336	20,874	15.62	10	.32	.22	.18
Newspaper		1,314	25,992	19.78	10	.30	.22	.22
Essays		1,135	26,349	23.21	10	.33	.21	.21
Fiction		2,753	31,491	11.44	10	.35	.22	.22
Travel-guides		919	17,154	18.67	10	.28	.23	.23
SemEval		1,192	7,760	6.51	10	.37	.20	.20
↓ *Fiction4*	1798-1965	6,300	73,250	11.6	>=2	.67	-	-
Hymns	1798-1873	2,026	12,798	6.3	2	.72	-	-
Fairy tales	1837-1847	772	18,597	24.1	3	.69	-	-
Prose	1952	1,923	30,279	15.7	2	.63	-	-
Poetry	1965	1,579	11,576	7.3	3	.59	-	-

Table 5: Datasets with valence annotation. Valence was annotated on a sentence basis, so 'N annotations' indicates the number of sentences. 'N annotators' indicates the number of annotators reported per sentence. IRR per dataset and category is shown in α. Since *EmoBank* lacks unique annotator IDs, we cannot correlate individual annotators' scores. Therefore, we use Krippendorff's α measures agreement across V-A-D ratings per item in the full dataset and in subcategories. Only Emobank includes the full V-A-D annotation. Note that texts not indicated as Danish (flag) are all in English.

E Train to test dataset correlations

Visualizations of portability between datasets for valence. This figure is a visualization of Table 2.

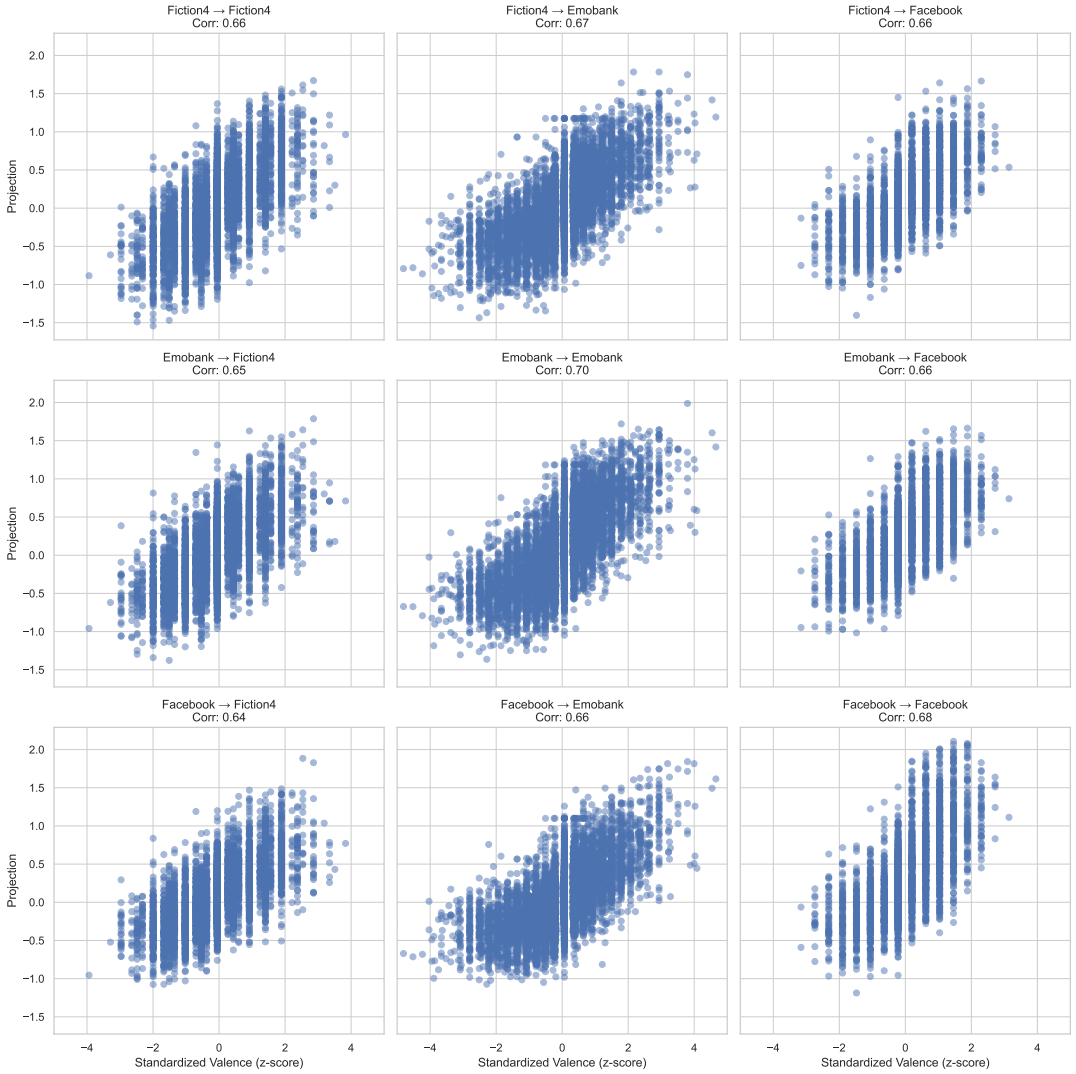

Figure 5: Relation between Concept Vector Projection scores (y-axis) and human scores (x-axis) on standardized valence across datasets. On top of each figure, the training set (on the left of the arrow) and the test set (on the right of the arrow) are shown.

F Beyond valence, visualized correlations

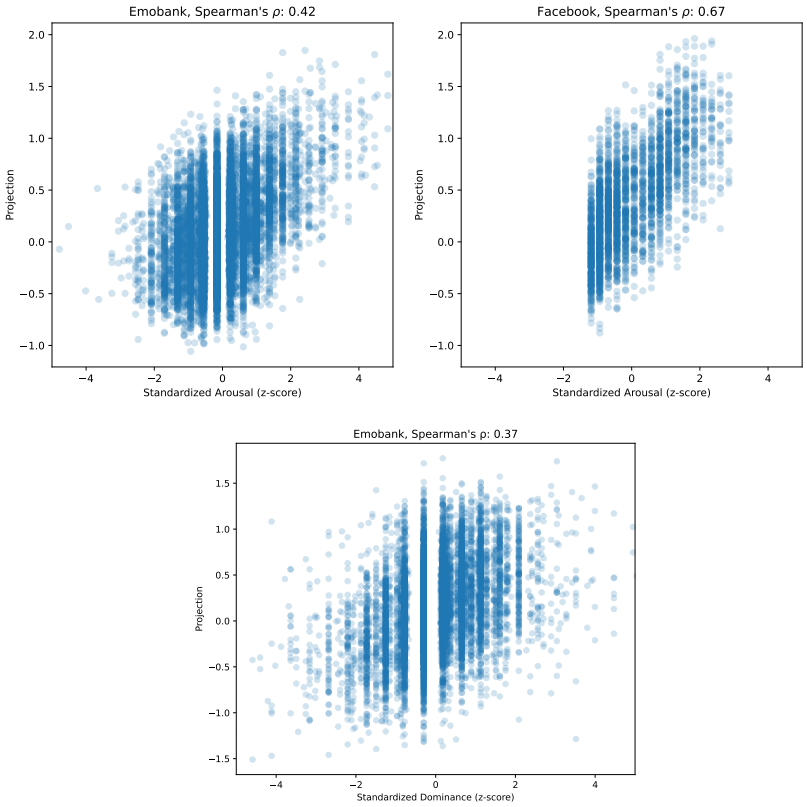

Figure 6: Top: Relation between Concept Vector Projection scores (y-axis) and human scores of standardized **arousal** in the Emobank and Facebook corpora. Bottom: Relation between Concept Vector Projection scores (y-axis) and human scores of standardized **dominance** in the Emobank corpus.

G Portability of Arousal

Visualizations of portability between datasets for arousal. This figure is a visualization of Table 2.

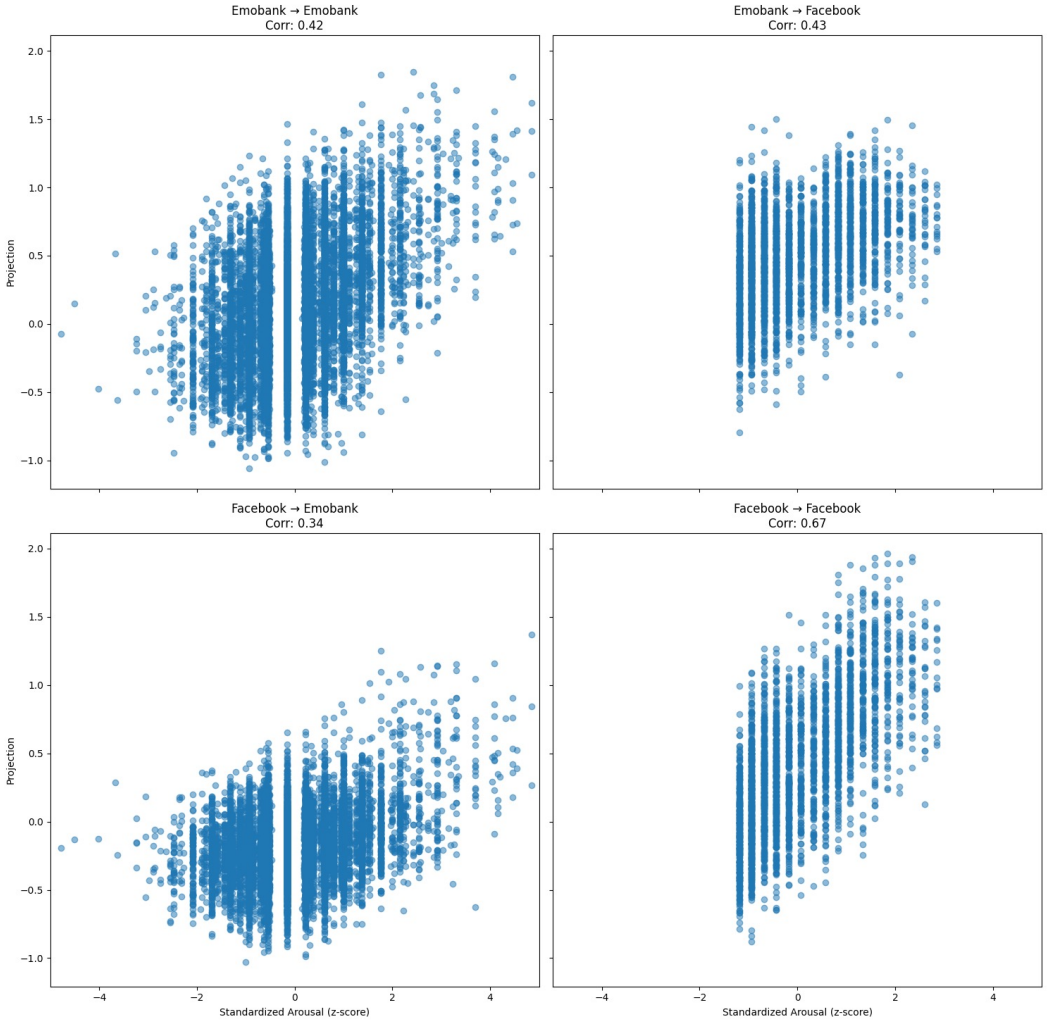

Figure 7: Relation between Concept Vector Projection scores (y-axis) and human scores (x-axis) on standardized arousal across datasets. A title such as Emobank→Facebook should be read as: Correlation between projections of arousal and human arousal ratings, when arousal vector is defined by the Emobank corpus and predictions are tested on the Facebook corpus.

H Downstream differences between Human Annotators and Projection models

As a sanity check on downstream analysis using projection scores instead of human annotators we tested a simple hypothesis. That both high and low valence scores correlate with high arousal. This would imply that arousal only correlates with valence, when we use the absolute value of valence (i.e. distance from the mean). While the slope of our linear regression varies between the two methods, we reach the same conclusion with both models. That there is a positive relation between absolute valence and arousal scores.

Figure 8: Scatterplots showing the relation between arousal and valence. The two top plots show no correlation between valence and arousal. The two bottom plots use absolute valence instead of valence, and depicts a positive significant relationship between absolute valence and arousal. Left side plots uses human annotations of EmoBank. Right side plots use projected ratings of EmoBank, and using the pos-neg vector defined on EmoBank.