Capturing the Affective Side of a Metaphor: a Comparison of BERT-based Sentiment and Prompted Large Language Models

Rebecca Guolo (rebecca.guolo@iusspavia.it)[1], Ginevra Martinelli[1], Chiara Barattieri di San Pietro[1], Valentina Bambini[1]

[1] *Laboratory of Neurolinguistics and Experimental Pragmatics (NEPLab), University School for Advanced Studies IUSS, Pavia, Italy*

Metaphors are a pragmatic phenomenon whose comprehension is guided by several psycholinguistic features. Among these, valence - framed within the Valence-Arousal-Dominance model [1] - plays a key role in shaping aesthetic, non-propositional effects [2,3]. Hence, human valence ratings are crucial to investigate the role of emotion in metaphor understanding. Yet, obtaining them is costly and time-consuming, motivating the search for computational alternatives. The aim of this work is to evaluate at what extent computational simulated values might approximate human valence judgments of metaphors. In particular, we evaluated (i) BERT-based sentiment analysis systems [4], (ii) LLM-based prompting [5], and (iii) a compositional strategy based on combining the valence of individual words [6].

As a starting point we focused on German as it offers, to the best of our knowledge, two datasets that include extended human affective ratings: the COMETA (n=60) [7] and the MIST (n=168) [8] datasets. Additionally, we included Italian everyday and literary metaphors from the Figurative Archive (n=996) [9], enriching it collecting ad-hoc human valence ratings (ICC=.72). For (i), sentiment ratings were derived from fine-tuned Italian [10] and German [11] BERT models. For (ii), GPT-4o was used because of its contextual processing ability. For (iii), valence scores for the topic and the vehicle of each Italian metaphor were extracted from the Multilingual Emotion Lexicon [12] and averaged to compute the metaphor's overall valence. Estimates were then analyzed via Pearson correlation analysis and via z-transformed Fisher's test to compare correlations' magnitudes.

BERT-based sentiment scores showed weak alignment with human ratings (Table 1): the German BERT performed poorly overall, while the Italian BERT showed moderate correlations (*rho=.598, p<0.001*), with different patterns between everyday and literary metaphors (*rho*=.498, *p*<0.001; *rho*=0.703, *p*<0.001, respectively), likely reflecting the greater affective load of literary language. Differently, GPT-derived ratings closely matched human evaluations (Table 1). Furthermore, GPT estimates captured human judgments of metaphor valence better than averaging the valence of the metaphor's constituent words (*Z*=-12.216, *p*<0.001 – Table 2). Although BERT-based sentiment scores capture human evaluation better than the compositional strategy (*Z*=5.897, *p*<0.001 – Table 2), GPT showed significantly better agreement with human values than BERT (*Z*=18.114, *p*<0.001 – Table 2).

Thus, our results suggest that metaphor valence goes beyond simple compositionality and is better estimated by models that capture contextual information.

References

[1] J. Russell. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178. https://doi.org/10.1037/h0077714

[2] N. Mashal & O. Itkes. (2014). The effects of emotional valence on hemispheric processing of metaphoric word pairs. *Laterality, 19*(5), 511-521. https://doi.org/10.1080/1357650X.2013.862539

[3] E. Ifantidou. (2021). Non-propositional effects in verbal communication: The case of metaphor. *Journal of Pragmatics, 181*, 6-16. https://doi.org/10.1016/j.pragma.2021.05.009

[4] D. Jurafsky & J. H. Martin. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models,* 3rd edition. Online manuscript released January 6, 2026. https://web.stanford.edu/~jurafsky/slp3/

[5] V. Mangiaterra, H. Al-Azary, C. Barattieri di San Pietro, P. Canal & V. Bambini. (2025). *Can GPT replace human raters? Validity and reliability of machine-generated norms for metaphors*. https://arxiv.org/abs/2512.12444

[6] S. M. Mohammad. (2025). *Breaking bad: Norms for valence, arousal, and dominance for over 10k English multiword expressions*. https://arxiv.org/abs/2511.19816

[7] F. M. M. Citron, M. Lee & N. Michaelis. (2020). Affective and psycholinguistic norms for German conceptual metaphors (COMETA). *Behavior Research Methods, 52*(3), 1056-1072. https://doi.org/10.3758/s13428-019-01300-7

[8] N. Müller, A. Nagels & C. Kauschke, C. (2022). Metaphorical expressions originating from human senses: Psycholinguistic and affective norms for German metaphors for internal state terms (MIST database). *Behaviour Research Methods, 54*, 365-377. https://doi.org/10.3758/s13428-021-01639-w

[9] M. Bressler, V. Mangiaterra, P. Canal, F. Frau, F. Luciani, B. Scalingi, C. Barattieri di San Pietro, C. Battaglini, C. Pompei, F. Romeo, L. Bischetti & V. Bambini. (2026). Figurative Archive: an open dataset and web-based application for the study of metaphor. *Scientific Data, 13*, 151. https://doi.org/10.1038/s41597-025-06459-7

[10] F. Bianchi, D. Nozza & D. Hovy. (2021). Feel-it: Emotion and sentiment classification for the Italian language. *In Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 76-83). Association for Computational Linguistics. https://aclanthology.org/2021.wassa-1.8/

[11] O. Guhr, A.K. Schumann, F. Bahrmann & H. J. Böhme. (2020). Training a broad-coverage German sentiment classification model for dialog systems. *In Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1620-1625). European Language Resources Association. https://aclanthology.org/2020.lrec-1.202/

[12] S. Buechel, S. Rucker & U. Hahn. (2020). Learning and evaluating emotion lexicons for 91 languages. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1202-1217). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.112/

Figures and Tables

Table 1: Spearman correlations between human ratings and BERT and GPT evaluations

Dataset	BERT	GPT
Figurative Archive	0.598***	0.893***
Everyday Metaphors	0.498***	0.868***
Literary Metaphors	0.703***	0.929***
German Datasets	0.158*	0.936***
COMETA	0.126	0.945***
MIST	0.146	0.937***

Table 2: Pairwise Fisher z-tests comparing the magnitude of the correlations between the lexicon-based compositional model, BERT, and GPT.

Comparison	z1	z2	z observed
mLemmas vs. GPT	0.824	1.372	-12.1216***
mLemmas vs. BERT	0.824	0.559	5.897***
GPT vs. BERT	1.372	0.559	18.114***

Optional Supplemental Information

English translation of the prompt given to GPT: "*You will be given a metaphor and asked to evaluate it based on its emotional valence. Emotional valence refers to how positive or negative the emotions elicited by the metaphor are. Use a rating scale ranging from 1 ("very negative") to 7 ("very positive"). If the metaphor elicits a very negative emotion, such as sadness or anger, then assign a score of 1. If the metaphor elicits a very positive emotion, such as happiness or love, assign a score of 7. Respond with only a number from 1 to 7; do not add anything else. Metaphor: met*"