From Sentiment to Valence in Metaphor: a Comparison of BERT-based Sentiment and Prompted Large Language Models

Rebecca Guolo, Ginevra Martinelli, Chiara Barattieri di San Pietro, Valentina Bambini
Laboratory of Neurolinguistics and Experimental Pragmatics (NEPLab)
University School for Advanced Studies IUSS, Pavia, Italy
rebecca.guolo@iusspavia.it, ginevra.martinelli@iusspavia.it,
chiara.barattieridisanpietro@iusspavia.it, valentina.bambini@iusspavia.it

Abstract

Although the affective dimension is a key aspect of metaphor, computational studies of figurative language have largely overlooked psycholinguistic variables such as valence. This study investigates whether computational models can reliably estimate the affective aspects of Italian and German metaphors and whether metaphor valence is compositionally derived. Outputs of BERT-based sentiment analysis and a valence-prompted LLM were compared with human ratings. Results show that the former exhibit limited alignment with human judgments, whereas higher agreement is achieved when the explicit concept of valence is prompted in a LLM. Both humans and models rely on the combined valence of the individual lemmas, suggesting a compositional contribution to metaphor valence.

Keywords: Valence, Metaphor, BERT, GPT

1. Introduction

Metaphors represent a complex pragmatic phenomenon that involves a gap between the literal and the intended meaning, which is typically resolved via context-based inferences (Grice, 1975; Searle, 1979). Among the psycholinguistic features known to guide metaphor comprehension - e.g., familiarity, concreteness, and imageability - the affective dimension appears to play a particularly prominent role in deriving the meaning of metaphors and producing aesthetic non-propositional effects (Blasko and Briihl, 1997; Montefinese et al., 2025; Gerwien et al., 2024; Ifantidou, 2021). Affective properties are commonly conceptualized within the Valence–Arousal–Dominance (VAD) framework (Russell, 1980), which provides a fine-grained model for operationalizing affect estimation. Within this framework, valence, defined as the degree of pleasantness or unpleasantness evoked by a stimulus (Jurafsky and Martin, 2026), has been shown to interact with cognitive processes underlying metaphor comprehension (Mashal and Itkes, 2014). The availability of reliable affective ratings is therefore crucial to investigate the role of emotion in metaphor understanding. However, collecting such ratings through human annotation is both time- and resource-intensive. For other psycholinguistic features, such as concreteness and familiarity, recent research has demonstrated that computational approaches, particularly Large Language Models (LLMs), can effectively approximate human ratings (Mangiaterra et al., 2025). To our knowledge, such methods have not yet been systematically applied to generate valence ratings for metaphorical expressions. To address this gap, this study aims to evaluate computationally generated valence ratings for metaphorical expressions using three computational approaches: (i) sentiment analysis systems (Jurafsky and Martin, 2026), which provide coarse-grained polarity judgments but may overlook subtle emotional nuances; (ii) LLM-based prompting, which elicits valence judgments by leveraging the capabilities of LLMs to process the language of emotion (Martínez et al., 2025; Conde et al., 2026) and figurative language (Fuoli et al., 2026); (iii) lexicon-based estimation, deriving valence scores from established word-level affective norms and aggregating them, allowing to test whether the valence of a metaphorical expression can be approximated from the valence of its constituent words (Mohammad, 2025).

2. Materials and Method

We collected a set of Italian and German metaphors from three different datasets: 1) the Figurative Archive dataset (Bressler et al., 2026), which includes 996 Italian metaphors differing in structure and semantic domains and divided into everyday (i.e., metaphors that occur in ordinary language, pooled from nine studies) and literary metaphors (i.e., sourced from poetry or prose). Specifically, metaphors were identified i)using keywords belonging to semantic classes known to be productive for metaphorical use, and; ii) searching for expressions in the “A of B” form; the resulting metaphors were then reviewed and problematic items removed. 2) the COMETA (Citron et al., 2020) and 3) the MIST (Müller et al., 2022) datasets, which collect respectively 60 conceptual German metaphors with a sentence-level structure, and 168 German metaphors involving human senses. While for the German datasets human affective ratings were al-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 212–216
16 May 2026.

ready available, the Italian metaphors were rated ad-hoc for this study by 9 PhD candidates in linguistics, all Italian native speakers (mean age = 26,8 years, SD = 1,5; 7 women) on a -3 +3 range. The average ICC was 0.72, indicating moderate inter-rater agreement.

Ratings of valence were derived from two BERT-based models (one for Italian, and one for German) and one GPT model. Valence ratings for Italian metaphors were obtained using *feel-it-italian-emotion model* (Bianchi et al., 2021). Valence ratings for German metaphors were derived from *germansentiment* (Guhr et al., 2020). Valence ratings from GPT were obtained from GPT 4-o (OpenAI, 2024) via API. We chose this model considering its capacity to process contextualized and semantically complex expressions. The Italian and German prompts[1] instructed the model to assess metaphor valence on a seven-point scale (same span given to humans). No examples were inserted in the prompt, applying a zero-shot learning technique. Spearman correlation analysis was applied to test the relationship between human ratings and GPT- and BERT-based ratings. Finally, to assess whether metaphor valence is compositionally derived, we extracted the valence for each Italian metaphor's topic (i.e., the subject of the metaphor) and vehicle (i.e., the term used to convey the metaphorical meaning) from the Multilingual Emotion Lexicon (MEL - Buechel et al. (2020)) and correlated them with the metaphor valence ratings provided by humans and models. Additionally, we computed the average valence of the lemmas composing each metaphor (excluding stopwords) and likewise correlated them with metaphor valence ratings obtained from humans and models. To compare the magnitude of the correlations for compositionality, Fisher's z-transformation was applied following Pearson correlation analyses. This procedure could only be conducted for the Italian metaphors, as the German datasets lack topic and vehicle annotations, preventing the extraction of MEL-based values.

3. Results

BERT models. Results show a moderate correlation between Italian BERT-based ratings and human ratings (*rho*=0.598, $p<0.001$), with different patterns between everyday and literary metaphors: while valence scoring for everyday metaphors show a moderate correlation (*rho*=0.498, $p<0.001$), BERT-based and human ratings for literary metaphors exhibit a strong correlation (*rho*=0.703, $p<0.001$). In contrast, we observe very weak and non significant correlations between human and German BERT-based ratings (*rho*=0.158, $p=0.017$), both for the COMETA and for the MIST datasets (respectively, *rho*=0.126, $p=0.337$; and *rho*=0.146, $p=0.058$).

GPT model. Findings are homogeneous across languages, with very strong correlations between GPT-based and human ratings both for Italian and German (*rho*=0.893, $p<0.001$; *rho*=0.936, $p<0.001$, respectively), and similar patterns for both everyday and literary metaphors in the Italian dataset (*rho*=0.868, $p<0.001$ and *rho*=0.929, $p<0.001$, respectively) and for both COMETA and MIST German datasets (*rho*=0.945, $p<0.001$ and *rho*=0.937, $p<0.001$, respectively).

Compositionality. Correlation analyses on human ratings (MEL) show moderate correlations between the valence of the topic and the overall metaphor valence (*rho*=0.604, $p<0.001$), as well as between the valence of the vehicle and the overall metaphor valence (*rho*=0.433, $p<0.001$). The mean valence of the topic and the vehicle composing the metaphor exhibits a strong correlation (*rho*=0.681, $p<0.001$) with the overall metaphor valence. Valence ratings derived from the Italian BERT-based model show moderate correlations with topic, vehicle and their mean valence (*rho*=0.480, $p<0.001$; *rho*=0.419, $p<0.001$; *rho*=0.592, $p<0.001$ respectively). Conversely, GPT-derived measures of valence show a strong correlation with the mean valence of topics and vehicle (*rho*=0.736, $p<0.001$), while correlations between topic and metaphor valence (*rho*=0.608, $p<0.001$), and between vehicle and metaphor valence (*rho*=0.513, $p< 0.001$) remain moderate.

Correlations' magnitudes for compositionality. The comparison between the mean valence (MEL) and GPT revealed a significant difference (z=-12.216, $p<0.001$), with GPT showing a higher correlation with human ratings (z=1.372) than the mean valence (z=0.824). Similarly, the comparison between the mean valence and BERT was significant (z=5.897, $p0<.001$), indicating a higher correlation for the mean valence (z=0.824) compared to BERT (z= 0.559). Finally, GPT and BERT also differed significantly (z=18.114, $p0<.001$), with GPT showing a substantially higher correlation with human ratings (z=1.372) than BERT (z=0.559).

[1] English translation of the given prompt: *"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"*

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 1: Correlation results between the BERT and GPT models for Italian and German ratings.

Category	Human	BERT	GPT
Topic	0.604***	0.480***	0.608***
Vehicle	0.433***	0.419***	0.513***
mLemmas	0.681***	0.592***	0.736***

Table 2: Correlation results with MEL data.

4. Discussion

While the validity of machine-generated single-word ratings has been widely established (Martínez et al., 2025; Conde et al., 2026), the present study aimed at assessing the validity of valence ratings for metaphorical expressions as derived from BERT and GPT, evaluating them against human ratings. Firstly, our results revealed that across languages (Italian and German), BERT-based sentiment ratings were weakly aligned with human ratings, whereas GPT derived ratings well approximated human judgment. In particular, the German BERT performed poorly, while the Italian BERT showed moderate correlations for everyday and strong for literary metaphors, likely reflecting the greater affective load of literary language. It cannot be however excluded that GPT-4o might have been exposed to parts of the datasets during training. This concern could apply primarily to the German datasets, as the Italian dataset was released after the model's training cutoff. If this was the case, GPT performance in German is only slightly higher compared to Italian. So, if prior exposure were the primary driver, a substantially larger gap would be expected. Instead, the similar performance across German and Italian suggests that the results are better explained by generalizable inferential processes rather than memorization. Secondly, our results on the mechanisms underlying metaphor valence indicate that, for both humans and models, the mean valence of the constituent words largely guides the overall evaluation. For GPT, the correlation with the mean lemma valence exceeds 0.70, suggesting a strong reliance on the valence of individual words. For both humans and models, the topic plays a central role compared to the vehicle, anchoring the metaphor's affective value. Overall, both BERT and GPT are better in capturing the evaluation of a metaphor's human valence than a compositionality model based on the average valence value of the words that compose the metaphor. More specifically, BERT-based sentiment analyzers do not reliably approximate human valence ratings for metaphors, whereas prompted LLMs closely align with human-like judgments.

Criterion: Human			
Comparison	**z1**	**z2**	**z observed**
mLemmas-GPT	0.824	1.372	-12.216***
mLemmas-BERT	0.824	0.559	5.897***
GPT-BERT	1.372	0.559	18.114***

Table 3: Correlations' magnitudes for compositionality.

5. Limitations

This work is a preliminary study and several limitations are worth noting. First, having few annotators might limit the generalizability of the findings, and future research should both increase their number and their heterogeneity in terms of sociodemographic features, also exploiting online surveys.

Second, the use of existing models and GPT-4o for valence prediction provides limited technical novelty. To address this, future studies could explore additional models, including open-source alternatives such as Llama, in order to obtain more robust and reliable results.

Third, we acknowledge the imbalance in the number of metaphors across the two languages, which should be accounted for in the statistical modeling.

Finally, the human Italian ratings achieved an ICC < 0.75, that is considered only moderate. From another perspective, however, this difference in annotation reliability from Italian and German datasets may also help explain the pattern observed with BERT. In particular, BERT shows better performance on the Italian data, where ratings are noisier, and lower performance on the German data, where ratings are more consistent. Instead, GPT performances are significant in both languages, suggesting that its predictions are less affected by annotation variability. This pattern indicates that GPT may capture subtle semantic properties of metaphors in a way that is closely aligned with human judgments, even when those judgments are based on a relatively small number of annotators.

6. Ethics

The present study aims to use computational tools to approximate human valence judgments in metaphorical language. However, we are aware that the used datasets do not cover all possible syntactic structures and semantic domains of metaphors and most importantly, they may reflect

linguistic and cultural bias. Therefore, human participants remain the gold standard for affective evaluation, while computational approaches must be considered as complementary tools rather than complete replacements.

7. Bibliographical References

F. Bianchi, D. Nozza, and 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*. Association for Computational Linguistics.

D. G. Blasko and D.S. Briihl. 1997. Reading and recall of metaphorical sentences: Effects of familiarity and context. *Metaphor and Symbol*, 12(4):261–285.

S. Buechel, S. Rucker, and U. Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.

J. Conde, G. Martínez, M. Grandury, C. Arriaga, J. Haro, S. Schroeder, F. Hintz, P. Reviriego, and M. Brysbaert. 2026. Updating the german psycholinguistic word toolbox with ai-generated estimates of concreteness, valence, arousal, age of acquisition, and familiarity. *Journal of cognition*, 9(1).

M. Fuoli, W. Huang, J. Littlemore, S. Turner, and E. Wilding. 2026. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning.

J. Gerwien, M. Filip, and F. Smolík. 2024. Noun imageability and the processing of sensory-based information. *Quarterly journal of experimental psychology*, 77(10):2137–2150.

H. P. Grice. 1975. Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.

O. Guhr, A. K. Schumann, F. Bahrmann, and 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*, pages 1620–1625, Marseille, France. European Language Resources Association.

E. Ifantidou. 2021. Non-propositional effects in verbal communication: The case of metaphor. *Journal of Pragmatics*, 181:6–16.

D. Jurafsky and 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.

V. Mangiaterra, H. Al-Azary, C. Barattieri di San Pietro, P. Canal, and V. Bambini. 2025. Can gpt replace human raters? validity and reliability of machine-generated norms for metaphors. Preprint: 2512.12444.

G. Martínez, J. Conde, P. Reviriego, and M.Brysbaert. 2025. Ai-generated estimates of familiarity, concreteness, valence, and arousal for over 100,000 spanish words. *Quarterly journal of experimental psychology*, 78(10):2272–2283.

N. Mashal and O. Itkes. 2014. The effects of emotional valence on hemispheric processing of metaphoric word pairs. *Laterality*, 19(5):511–521.

S. M. Mohammad. 2025. Breaking bad: Norms for valence, arousal, and dominance for over 10k english multiword expressions. *arXiv preprint arXiv:2511.19816*.

M. Montefinese, A. Visalli, A. Angrilli, and E. Ambrosini. 2025. Fine-grained concreteness effects on word processing and representation across three tasks: An erp study. *Psychophysiology*, 62(5).

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

J. Russell. 1980. A circumplex model of affect. *Journal of Personality and Socialy Psychology*, 39:1161–1178.

JR. Searle. 1979. *Expression and Meaning*. Cambridge: Cambridge University Press.

8. Language Resource References

M. Bressler and V. Mangiaterra and P. Canal and F. Frau and F. Luciani and B. Scalingi and C. Barattieri di San Pietro and C. Battaglini, C. Pompei and F. Romeo and L. Bischetti and V. Bambini. 2026. *Figurative Archive: an open dataset and web-based application for the study of metaphor*. Scientific Data, 151.

F.M.M. Citron and M. Lee and N. Michaelis. 2020. *Affective and psycholinguistic norms for German conceptual metaphors (COMETA)*. Behavior Research Methods, 52(3).

N. Müller and A. Nagels and C. Kauschke. 2022. *Metaphorical expressions originating from human senses: Psycholinguistic and affective norms for German metaphors for internal state terms (MIST database)*. Behaviour Research Methods, 54.