Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond

Pascale Feldkamp[a], Ea Lindhardt Overgaard[b], Kristoffer Nielbo[a] and Yuri Bizzoni[a]

[a] *Center for Humanities Computing Aarhus, Jens Chr. Skous Vej 4, Building 1483, 8000 Aarhus C, Denmark*
[b] *School of Communication and Culture, Jens Chr. Skous Vej 2, Building 1485, 8000 Aarhus C, Denmark*

Abstract

As they represent one of the most complex forms of expression, literary texts continue to challenge Sentiment Analysis (SA) tools, often developed for other domains. At the same time, SA is becoming an increasingly central method in literary analysis itself, which raises the question of what are the challenges inherent to literary SA. We address this question by probing units from a variety of literary fiction texts where humans and systems diverge in their valence scoring, seeking to relate such disagreements to semantic traits central to implicit sentiment evocation in literary theory. The contribution of this study is twofold. First, we present a corpus of valence-annotated fiction – English and Danish language literary texts from the 19th and 20th centuries – representing different genres. We then test whether sentences where humans and models disagree in sentiment annotation are characterized by specific semantic traits by looking at their distribution and correlation across four different corpora. We find that items where humans detected significant sentiment, but where models did not, consistently employ lower levels of *arousal*, *dominance* and *interoception*, and higher levels of *concreteness*. Furthermore, we find that the amount of human-model disagreement correlated with semantic aspects is linked to the interiority-exteriority continuum more than with direct sensory information. Finally, we show that this interaction of features linked to implicit sentiment varies across textual domains. Our findings confirm that sentiment evocation exploits a more diverse and subtle set of semantic channels than those observed through simple sentiment analysis.

Keywords

sentiment expression, literary language, implicitness, objective correlative, sentiment analysis

1. Introduction

Sentiment Analysis (SA) is an increasingly central method for computational literary research [1], an especially popular application being that of gauging the 'sentiment arcs' of novels, i.e., the 'shapes of stories' [2, 3, 4]. Still, the relation between valence extracted with SA tools and the human perception of literary texts at a granular level remains an open question, as tools applied to the literary domain are primarily geared towards processing nonliterary texts.

While some recent work has examined the adequacy of available SA tools for literary analysis [5, 6, 7], the question of how to validate them – against whose judgements, and at what level (i.e., the story-level vs. sentence-level, etc.) – is a persistent concern, also due to the relative scarcity of annotated resources in the literary domain. Moreover, the observed inadequacy of SA tools for literature has raised the question of what the difference of literary texts might be

CHR 2024: Computational Humanities Research Conference, December 4-6, Aarhus, Denmark.
pascale.moreira@cc.au.dk (P. Feldkamp); @cc.au.dk (E.L. Overgaard); kln@cas.au.dk (K. Nielbo); yuri.bizzoni@cc.au.dk (Y. Bizzoni)
0000-0002-2434-4268 (P. Feldkamp); 0000-0000-0000-0000 (E.L. Overgaard); 0000-0002-5116-5070 (K. Nielbo); 0000-0002-6981-7903 (Y. Bizzoni)

in comparison to the nonliterary [8], where tools seem to perform comparably better [9]. How does the "literary" differ in its way of communicating sentiments to readers?

In fact, due to their textual complexity, literary texts are often said to differ from more *communicative* texts [10]: they are effective at multiple narrative levels [11, 12]; creatively divergent from standard language use [13, 14]; reliant on poetic devices [15]; and ambiguity, effecting contesting interpretations [16]. Moreover, literary language may exhibit various strategies for conveying emotion beyond simply using words directly associated with emotional states (e.g., "sad"). While language on, e.g., social media, may also rely on omission and subtlety, literary theorists have frequently claimed that literariness or the *poetic* function of language excels in this regard and is distinct from its more directly communicative function.[1] Regarding sentiment expression, it has recently been suggested that literary prose relies on specific semantic traits connected to affective understatement and concreteness to *evoke* – rather than "communicate" – sentiment in readers [8].[2] These features are especially interesting since they are related to the seminal concept of the "objective correlative" in literary theory, which suggests that literary texts effectively convey emotions by grounding them in – concrete and objective, rather than emotional and subjective – entities and situations [20], while avoiding abstract and emotional language.[3] In this study, we further pursue the hypothesis that literary texts rely on omission and materiality, i.e., features connected to the concept of the objective correlative, to evoke sentiments, rather than simply mention them.

While divergences between human and model judgments in sentiment analysis are generally taken to indicate shortcomings in SA tools, they present an interesting case for testing the difference of literary language in expressing sentiment, assuming that SA tools are generally tuned toward more explicit forms of communication due to their development on e.g. social media – also suggested by more general studies of implicitness [22, 23]. Using a single novel as their data, Bizzoni and Feldkamp [8] explored this discrepancy, finding that certain semantic elements – levels of *arousal*, *dominance* and *concreteness* – are indicative of human/model disagreement. To gain further insight into sentiment expression in literary texts, we go one step further: we test the distinguishing power of these features in a much larger corpus of annotated literary texts. We also extend the list of features by an additional four: *imageability*, *visuality*, *hapticity*, and *interoception*, in order to further examine the connection of human/model disagreement with features that relate to the concept of the objective correlative and of omissive strategies in literary sentiment expression.[4]

We conduct two experiments. First, we test whether sentences where models and humans agree can be distinguished from sentences where they disagree based on these specific semantic traits. Secondly, we further explore the robustness of the relation between features and sentiment *evocation* by testing the correlation of these features to the absolute difference in sentiment scoring between humans and models.

[1] Jakobson distinguished the "poetic function" of language from its "emotive or expressive function", which "aims a direct expression of the speaker's attitude toward what he is speaking about" [10, p. 66]

[2] Again, these phenomena naturally extend outside the literary domain [17]: tweets using irony or figurative language, e.g., likely effect diverging reader interpretations [18, 19].

[3] Besides T.S. Eliot, who coined the term, other famous proponents of this view are the Imagists [21].

[4] All data and code for the present study is available here.

2. Related works

2.1. Literary sentiment analysis

While SA tools perform increasingly well in some domains [24, 9], studies have pointed out the cross-domain drop in performance, as well as the lagging behind of tools for under-resourced languages [25, 26, 27]. Still, some studies have suggested that Transformer-based models might be able to bridge the gap and perform better on literary or poetic material [28]. Assessing the performance of models on historical Danish and Norwegian literary texts, Allaith et al. [29] found that multilingual Transformers outperformed both fine-tuned models and classifiers based on lexical resources in the target language, which aligns with the findings of Schmidt et al. [28] and Schmidt and Burghardt [30] for historical German drama. Still, the disparity between human and model sentiment judgements in literary texts continues to be observed [6, 7]. The disparity is often related to the effects of narrative, annotators generally having access to the narrative context of a sentence, rather than to differing strategies in sentiment expression across domains.

2.2. Literary sentiment expression

The concept of "implicit" expression is particularly relevant, and complex, in literary writing. Several theories of writing point to the importance of avoiding concepts or ideas (however this is intended) in a too explicit way. The widely known precept of "Show Don't Tell" points at least partly in this direction [12]. Moreover, critics continually rely on terms like emotional "evocativeness" and "understatement" to describe writing styles [31, 32]. Literary and affect theory has also more recently foregrounded the use of materiality and sensuousness in literature – including poetry – to evoke affective reactions in readers, emphasizing the way objects and things are culturally invested with meaning and affect [33, 34] and thus utilized by authors to evoke embodied, affective experiences [35].[5] Despite the significance of implicit, evocative, and expressive strategies, there is little consensus on how to reliably track these in literature and whether such types of expression have recognizable linguistic markers.

The association of materiality to literary evocation is not new, and closely relates to the Modernists' and New Critics' valuation of *concreteness* over *abstraction* [31], as well as to the idea of the 'objective correlative' which T.S. Eliot [20] proposed in 1948. Eliot suggested that the effective way of expressing emotion in literature is 'by proxy', through an external and objective – in the sense of intersubjectively recognizeable – "set of objects, a situation, a chain of events which shall be the formula of [a] particular emotion" [20]. The concept of the objective correlative suggests that literary language effectively evokes sentiments in readers by being both *more omissive* (relying less on directly emotion-associated words)[6] and *more concrete* (relying on objects and situations).

This hypothesis has been supported by computational literary studies. Auracher and Bosch [38] found that the concreteness of literary language impacts the emotional engagement of readers and their experience of suspense, and Bizzoni and Feldkamp [8] tracked the "omissive" writing of Ernest Hemingway by looking at the amount and intensity of sentiment expressions

[5] In narratology, this is close to what Fludernik has termed narrative "experentiality" [36]. Burroway [37] explicitly notes that when using of nouns that evoke sense images and verbs that represent visualizeable actions "the writing comes alive".

[6] The advice against "sentimentalism" and abstraction in literary language is present in Eliot, though more prominent in e.g., Ezra Pounds' literary criticism [21].

Table 1
Used datasets with valence annotation. For all but the *FB* dataset, valence was annotated on a sentence basis, so the number of annotations generally indicates a number of sentences. The Spearman correlation (ρ) between the human mean and RoBERTa scores (H/R) is provided (for all, $p < .01$). Summing up, the total number of annotated lines considered in this study is $n = 19,327$. The Annotator/line indicates the number of annotators of valence reported per line in the corpus.

	N. annotations	N. words	$\bar{x}$ words/line	Annotators/line	H/R correlation
FB	2,895	57,436	19.8	2	0.78
EmoBank	8,735	173,958	19.9	10	0.65
Letters	1,344	25,550	19.0	10	0.69
Blog	1,323	25,691	19.4	10	0.68
Newspaper	1,308	31,647	24.2	10	0.62
Essays	1,131	30,958	27.4	10	0.62
Fiction	2,711	39,393	14.5	10	0.58
Travel-guides	918	20,719	22.6	10	0.48
Fiction4	6,300	73,250	11.6	>2	0.64
Hymns	2,026	12,798	6.3	2	0.67
Fairy tales	772	18,597	24.1	3	0.57
Prose	1,923	30,279	15.7	2	0.59
Poetry	1,579	11,576	7.3	3	0.56

detectable in sentences, compared to how "expressive" (in terms of sentiment) readers perceive these sentences to be. Comparing sentences where humans and models agree vs. those where they disagree, they found that arousal and dominance levels (of the NRC-VAD lexicon [39]) were indicative of omissive strategies of evocation. Moreover, the level of concreteness of the language used appeared higher in sentences with higher disagreement between humans' and models' valence attribution.

2.3. Semantic traits of literariness

In computational linguistics, the use of semantic traits – often derived from psycholinguistic norms – for the study of narrative is relatively frequent [40, 41, 42]. Yet, their use to model poetic literary strategies of sentiment evocation is still relatively pioneering. Kao and Jurafsky [41] applied semantic measures to poetry: imageability to gauge "imagery"and concreteness to gauge "concrete imagery", using the concreteness ratings of Brysbaert et al. [43], along with objective/abstract word categories [44], as well as psycholinguistic norms [45] to model "emotional language". They show that these features differ between "amateur" and appraised poets, where appraised poets use more concrete and imageable language and less emotional words. Conversely, Maslej et al. [42] find that abstraction and arousal correlate positively with the perceived interest of readers in fictional characters, which is perhaps related to the general tendency of abstract concepts to be more emotionally valenced than concrete ones [46]. Ullrich et al. [40] observe how the differences between perceived affect in poem annotations can be largely explained by lexical psycholinguistic norms, with sentiment dimensions like arousal being one of the best predictors of the perceived affective meaning of the poems. The studies show above all that imageability, concreteness and arousal strongly relate to how readers feel about literary texts.

3. Data

3.1. Datasets

To probe textual features of implicit sentiment, we created a diverse corpus of fiction spanning four genres, manually annotated for valence (*Fiction4*). For the second experiment of this paper – comparing more or less literary genres – we also selected other datasets to represent a diversity of genres in the literary and non-literary domains which had all been annotated for valence on a continuous scale.[7] As such, we include social media, journalism, and genres with various degrees of literariness (essays, letters, travel-writing, fairytales), which may also be thought to represent degrees of colloquialism (from blogs to journalism).[8]

***Fiction4*:** The new dataset presented in this study, with human annotations for valence ($n = 6,300$).[9] It includes four different genres – fairy tales, hymns, prose and poetry – over the 19$^{\text{th}}$ and 20$^{\text{th}}$ century, in one high- and one low-resource language (English and Danish). The corpus was compiled with an aim toward diversity (in genre, time and language) while still aiming to include well-known and culturally significant works by both male and female authors.[10] We also took into account both i) the texts' cultural significance and ii) their narrative and poetic complexity, which may represent a particular challenge to SA tools. The authors selected were Ernest Hemingway, Sylvia Plath, H.C. Andersen and the various authors of official hymn-books (for details, see Appendix, Table 7).

i) Regarding their cultural significance, for the English texts, Hemingway is perhaps one of the most famous 20$^{\text{th}}$ century authors for his prose, and his texts are read in education[11] and among the public.[12] Plath, similarly, is a widely read and acclaimed poet, perhaps the best known female American poet of the 20$^{\text{th}}$ century.[13]

For the Danish texts, Andersen's production is arguably the most central in Danish literary heritage [48]. While being less known internationally, the official hymnal book is the most widely distributed "poetry book" in Denmark [49],[14] used in the Danish education system at all levels, and shapes national cultural identity [50].

ii) Regarding the complexity of the corpus for the sentiment annotation task, we ensured variance across genre, place and time, but also emphasized, from a literary perspective, the level of literary complexity, including texts that could be considered either particularly challenging or particularly simple. We consider Hemingway's *The Old Man and the Sea* and Plath's *Ariel* as two ideally *difficult* cases for testing SA tools. Hemingway is known for an especially "omissive" writing style, direct and limited in its use of figurative language [51], while relying on implication rather than "overt emotional display", leaving much inference up to the reader [31]. Hemingway's *The Old Man and the Sea* (1952) has been considered emblematic for this

[7] We have adopted a very broad understanding of genre for this paper, encompassing more or less literary genres.

[8] We standardized the varying valence scales of chosen corpora to a scale from -1 to 1 (negative to positive).

[9] For an overview table of this corpus, see Appendix.

[10] Note, however, that the corpus is highly skewed towards male writers, not least because of the time-period and genres covered (i.e., hymns). For a detailed overview of the corpus, see Table 7 in Appendix.

[11] *The Old Man and the Sea* being studied in schools across the world [47].

[12] As of today, *The Old Man and the Sea* has over 1 million ratings on GoodReads.

[13] Plath's prose work *The Bell Jar* has around 1 million rations on GoodReads, and her poems appear among the top 250 assigned works on English Literature college syllabi.

[14] Note that the Danish term used "lyrik" encompasses poetry and songs.

minimalist style, which may omit characteristics that models rely on in sentiment scoring. Plath's poetry collection *Ariel* (1965) is complex in a slightly different way.[15] The so-called confessional poetry genre, of which Plath is considered emblematic, foregrounds idiosyncratic personal psychology and experiences against the "emotional vacuity of public language" and universal symbols [52]. In *Ariel* Plath writes on complex and political themes in idiosyncratic style consisting of "hallucinatory images" and novel metaphors [53], and the work has been used as a case of literature posing particular difficulty to most readers [54].

Conversely, we consider Andersen and religious hymns two ideally *simple* cases for testing SA tools. Andersen's fairy tales[16] are characterized by an essential simplicity, both stylistically an in their narrative progression [56, 57] and their ability to engage both children and adult readers [56]. Religious hymns[17] are characterized by their limited number of themes (e.g. worship, thanksgiving, etc.), which are are expressed through well-known and formalized (as well as recurring) phrases, metaphors, figurative and symbolic language [58]. The hymns' repetitive and predictable use of language may make them more accessible to models, even though their archaic and nuanced style may present challenges.

After collecting the texts, we found that some of these simplicity/complexity assumptions are reflected in the correlation between human and model valence scores – at least when using our method – where Plath's poetry shows the lowest correlation, and hymns the highest correlation (Table 1).

Reference corpora and datasets:

EmoBank The *EmoBank* is a multigenre corpus with human annotations for valence ($n = 8,735$),[18] with 10 annotators per sentence [59].[19] The corpus was composed from various categories in the *Manually Annotated Sub-Corpus of the American National Corpus* (MASC),[20] consisting of texts from 1990 and onwards [60]. We consider the *EmoBank* categories: Letters, Blog, Newspaper, Essays, Fiction, and Travel guides, which are relatively balanced (Table 1) including both longer and shorter texts within each category.[21]

FB The Facebook corpus of posts (FB)[22] collected between 2009 and 2011, consists of 2,895 status updates, each by a unique user, with human annotations for valence and arousal, with 2

[15] *Fiction4* includes all 40 poems in *Ariel*.

[16] *Fiction4* includes three of Andersen's most known fairy tales: "The Little Mermaid" (1837), "The Ugly Duckling" (1844), and "The Shadow" (1847), in an edition where spelling has been slightly modernized [55].

[17] *Fiction4* include 65 hymns from three different official hymnal books from the years 1798 ($n = 35$), 1857 ($n = 17$), and 1873 ($n = 13$). Years refer to publication years of three official church hymn collections, and hymns are collected at random.

[18] We exclude the heterogeneous SemEval category, as well as very short strings of sentences (noise) across the categories (length < 2).

[19] 5 annotators annotated each sentence for valence from a "reader" and "writer" perspective, i.e., 10 valence annotations per sentence. The valence scores represent a weighted average. See the documentation.

[20] Which is in turn a subset of the American National Corpus.

[21] The category 'essays', for example, comprises 8 texts, including the essay "A Brief History of Steel in Northeastern Ohio" or one on discrimination. 'Fiction' comprises 6 works of various genres, e.g., Richard Harding's "A Wasted Day" and the SciFi story "Captured Moments". Newspapers include various short reports (e.g. "A.L. Williams Corp. was merged into Primerica Corp." etc.) and longer reportages. Note that Travel Guides are generally written in a running prose, and includes both place-histories (e.g. "A brief history of Jerusalem") and current-day reflections (e.g. "Dublin and the Dubliners"). See the full MASC corpus here.

[22] https://github.com/wwbp/additional_data_sets/

annotators per post [61]. The *FB* dataset differs from our other corpora, consisting of posts – not sentences. While some posts are short (e.g., ":)" and "LOL"), the average length of posts is comparable to the average sentence length in, e.g., *EmoBank* (Table 1).

Beyond these corpora, we also include two datasets without valence annotation for comparison in terms of feature levels:

Image-captions ($n = 3,334,173$) of the Conceptual Captions dataset [62], of which we consider feature values to represent a "high-water mark" (i.e. high level) of language relying on object description and visuality.[23]

Participants free emotion event descriptions ($n = 6,898$)[24] of the International Survey on Emotion Antecedents and Reactions (ISEAR) [63],[25] of which we consider feature values to represent a "high-water mark" of language dealing with interiority (i.e., relating to inside sensations and self) and emotionality.

4. Methods

4.1. Human and automatic sentiment annotation

Model annotation Multilingual transformer-based models have shown best performance in SA for literary texts across languages (also for historical texts)[6, 29, 28] compared to dictionary-based approaches explicitly developed for literary texts as well as monolingual English models [6]. We, therefore, used the RoBERTa base xlm multilingual, finetuned for sentiment analysis on Twitter data,[26] which is comparable to the state-of-the-art models in a monolingual (non-literary) setting [64], and shows the best performance in the limited studies there are on literary prose [6].[27] XML-RoBERTa[28] was developed through a cross-lingual language training method, designed to boost its proficiency in comprehending and processing multiple languages by transferring skills it has acquired from one language to another.

With this model, we scored all sentences across our bilingual datasets.[29] The model returns polarity *positive* or *negative*, and a *neutral* label. To attain more continuous, nuanced data from the transformers' categorical output, we opted for using the same strategy as in Bizzoni and Feldkamp [6], i.e., using the confidence score of model labels as a proxy for sentiment intensity.

[23] https://github.com/google-research-datasets/conceptual-captions

[24] To exclude noise and non-answers (e.g., "I cannot remember"), we set an arbitrary threshold of 30 tokens for a description to be included, resulting an a diminished dataset from the original 7,659 datapoints.

[25] https://github.com/sinmaniphel/py_isear_dataset/

[26] We used this model off-the-shelf, so that the hyperparameters are as reported in: https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment/blob/main/config.json.

[27] Note that recent studies have tested newer, generative models for literary SA, notably Rebora et al. [7]. For this study, we excluded GPTs from our pool of tools. As our interest is not achieving top performance, but rather understanding the differences between SA tools and human annotation, we sought to employ only models that were designed for sentiment analysis and that don't depend on prompt engineering.

[28] https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta

[29] For the Danish texts, we tried the model both on the original Danish, on non-validated Google translations, and on manually checked and revised Google translations. We chose to use the model's output on validated translations, since the those valence scores correlated best with human annotations – the correlation with human mean values when applying the model on hymns and fairy tales in the original Danish was $\rho > .45,\ p < .01$, on Google-translations $\rho > .61,\ p < .01$, and on validated English translations $\rho > .65,\ p < .01$.

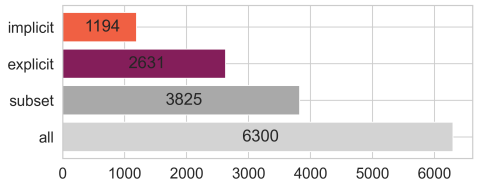


Figure 1: Groupsizes: the full *Fiction4*, the filtered subset (where human valence was below 4.5 and above 5.5 on the 0-10 scale), as well as the implicit and explicit groups. Number of sentences on the x-axis.

For example, a sentence with a *positive* label and a confidence of, e.g., 0.75, is interpreted as a valence score of +0.75. Similarly, a *negative* label with confidence of 0.89, is interpreted as a valence score of -0.89. For the neutral category, confidence is disregarded and levelled to a score of 0 (midscale or "neutral"). [30] The correlation between human mean score and the transformed RoBERTa score appears high across our selected corpora (Table 1).

As seen in Table 1, RoBERTa values correlate most strongly with human annotations of the *FB* dataset, while correlations with annotations of fiction (*EmoBank* and *Fiction4*) are much lower, possibly reflecting better development the model for certain more colloquial domains (social media, blogs, letters).

Human annotation of *Fiction4* Human annotators (at least $n = 2$/line) read the literary texts from beginning to end, scoring each line on a 0 to 10 valence scale:[31] 0 signifying the lowest, and ten the highest valence.[32]

The valence score was intended to represent the sentiment the sentence and verse expressed, and annotators were instructed to avoid rating how a sentence or verse made them feel and to try to report only on the sentiments embedded in the sentence, i.e., to think about the valence of the individual sentence and verse, without overthinking the story's/poetry's narrative.

It is worth noting that humans rarely reach an agreement higher than 80% (or 0.80 Krippendorff's α) for tasks like positive/neutral/negative discrete tagging [65] on *nonliterary texts* – and have lower agreement for continuous scale polarity annotation [66], especially for literary texts [7].

4.2. Sentence subsets

For our first experiment, exploring the prevalence of semantic traits in sentences where humans/model disagree and sentences where they agree, we divided our *Fiction4* corpus into two groups. First, we filtered out sentences which humans did not perceive any strong sentiment (i.e., with human valence scores between 4.5 and 5.5 on the 0-10 scale). On one hand, we then took sentences in which our chosen model did not assign any strong sentiment (below an

[30] Naturally, there are caveats to transforming sentiment polarity to continuous valence scores in this way. However, the approach has been shown to outperform dictionary-based (outputting continuous scores by design) and to approximate a human continuous valence annotation in literary prose [6]. Note that the distribution of transformed scores still tend to "look polar" as confidence score tend to be generally high, see Fig. 6, Appendix.

[31] "Lines" refer to sentences in the case of prose and to verse-lines in the case of the hymns/poetry. Sentences were tokenized using the nltk tokenize package.

[32] Annotators were researchers, three with a background in literary studies and one in cognitive science. The two annotators of the hymns (MA and PhD of literature) had domain knowledge in 19$^{\text{th}}$ century Scandinavian literature and historical religious hymns.

absolute score of 0.1, i.e., between -0.1 and +0.1)[33] and, on the other hand, sentences where it did. With this procedure, we distilled two groups of sentences (Fig. 1): one of sentences with humans/model disagreement, which we call the "implicit" group ($n = 1,194$) and one of human/model agreement, which we call the"explicit" group ($n = 2,631$).

4.3. Features

Based on previous work (section 2), we include three previously used semantic traits to examine their bearing on instances of implicit sentiment evocation: at the sentiment dimension, arousal,[34] and dominance,[35] and at the sensorimotor dimension, concreteness,[36] imageability, as well three additional sensory traits, visuality,[37] haptic,[38] and interoception.[39] We use the datasets below to measure sentence semantic trait values, averaging the score per feature for each sentence.

Concreteness lexicon: The lexicon by Brysbaert et al. [43] provides concreteness ratings for 37,058 English words. Annotators were recruited via Mturk (English native speakers). Each word was annotated by at least 25 annotators, on a scale from 1 (=most abstract, i.e., what cannot be experienced directly but the meaning of which is defined by other words) to 5 (=most concrete, i.e., what can be experience directly through one of the five senses). These ratings have been widely used [70] also in the literary domain [38].

NRC-VAD lexicon The lexicon by Mohammad [39] provides ratings of 20,000 English words on three sentiment dimensions (valence, arousal, dominance). Annotators were recruited via CrowdFlower, and each word was annotated by at least 6 annotators with a best/worst scaling approach (e.g. most arousal vs. least arousal). The lexicon has been used widely, as well as integrated in the SA tool VADER [71].

The Lancaster Sensorimotor Norms: The dataset provides norms of sensorimotor strength for 39,707 English words across 6 perceptual modalities (haptic, auditory, olfactory, gustatory, visual, and interoceptive)[69]. While the dataset includes action effectors (i.e., body parts) we used only the general perceptual norms, selecting only those we deemed especially relevant to the idea of objective correlative (i.e., material, visual objects/situations): visuality, interoception and hapticity. The perceptual part of the dataset had 2,625 annotators recruited via Mturk. Each word was rated from 0 (=not experienced with sense X) to 5 (=experienced greatly with sense X). These perceptual modality ratings have been used in, e.g., metaphor detection [72], and have served as a form of "embodied experience" information to enrich and improve language models [73].

Imageability The MRC Psycholinguistic Database (MRCPD) provides 26 linguistic and psycholinguistic variables for 150,837 English words – a subset of which are 9,240 words words

[33]Note that the model valences range from -1 to 1 (negative to positive), where 0 represents neutral.
[34]The degree to which a word prepares for action, captures or focuses attention [67].
[35]The degree of control evoked [68].
[36]The degree to which a word denotes a perceptible entity [43].
[37]The degree to which a word is experienced with the eyes [69].
[38]The degree to which a word is experienced by touch [69].
[39]The degree to which a word is experienced by sensations inside the body [69].

Table 2
Inter Rater Reliability between annotators across literary genres, using the mean ($\bar{x}$) of Spearman's ρ between pairs (for all, $p < .01$) – with Krippendorff's α for reference.

	Hymns	Fairy tales	Prose	Poetry
Spearman's ρ ($\bar{x}$)	0.73	0.68	0.62	0.59
Krippendorff's α	0.72	0.69	0.64	0.59

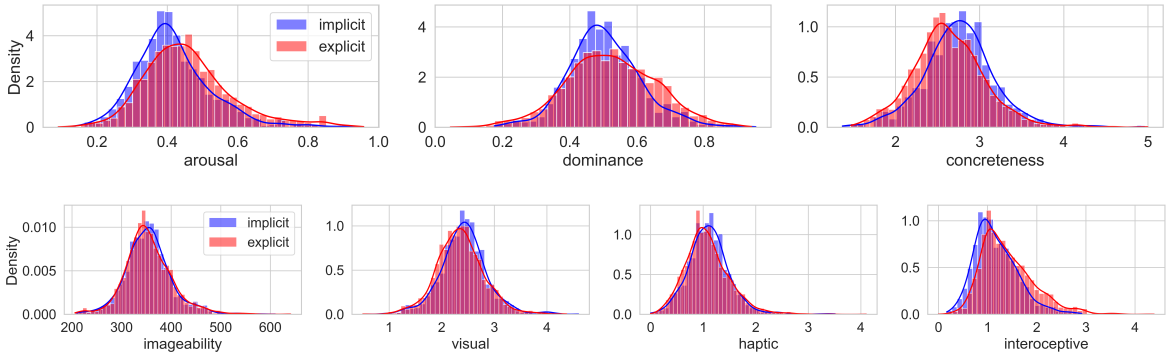


Figure 2: Distribution of feature values for the two groups: implicit and explicit groups of sentences in the *Fiction4* corpus.

rated for imageability [74]. These words have been rated by annotators. Ratings reflect *how easily a word can evoke mental imagery*, and fall in the range $100 - 700$. The lexicon has been used variously, e.g., in metaphor [75] and literary studies [41].

5. Results

5.1. Human annotation

We report a relatively high inter-rater reliability (IRR): between annotators, we find a mean correlation (Spearman's ρ) from 0.59 for poetry to 0.73 for hymns (Table 2).[40] IRR is high, especially for hymns, considering both the fragmentariness of the verses, and that humans tend to have low agreement for continuous-scale annotation (Section 4.1).

5.2. Experiment 1

For our first experiment we compared the two groups of sentences in *Fiction4* along each of the chosen features. We report the Mann-Whitney U-test effect's size and significance levels in Table 3.[41] We find that the strongest effect size is for interoceptive values (Table 3),

[40] As annotators operated within a continuous valence spectrum, divided into ten categories, we find that a correlation measure more clearly reflects direction and nuance of annotations (parallelity vs exactness), compared to categorical IRR measures. Therefore, we report Spearman's ρ and provide Krippendorff's α for reference (the level of measurement is considered *interval*).

[41] For the test, we dropped sentences with NaN-values in the specific feature we were testing, the number of dropouts was < 40 in each test.

Table 3
Mann-Whitney U test on the implicit vs. explicit group of sentences ($n = 1,194/2,631$) in the *Fiction4* corpus. Note that, for better readability, we have divided the effect size by 100 here. *$p < .01$, numbers in grey are $p > .05$.

	Concret.	Arousal	Dominance	Imag.	Visual	Haptic	Interocept.
Fiction4	1,278*	1,807*	1,636*	1,538	1,413*	1,461*	1,975*

Table 4
Mann-Whitney U test on the implicit vs. explicit group of sentences in the **reference corpora**. Again, the effect size has been divided by 100 for readability. *$p < .01$, numbers in grey are $p > .05$.

	Concret.	Arousal	Dominance	Imag.	Visual	Haptic	Interocept.
EmoBank	2,205*	3,320*	2,940*	2,429*	2,528*	2,745	3,380*
Letters	759*	1,142*	1,125*	797*	903	1,034	1,194*
Fiction	2,137*	2,959*	2,362	2,223	2,312*	2,348	2,865*
Blog	414*	658*	613*	475	473*	500	661*
Newspaper	464*	685*	591	542	547	607	762*
Essays	222	350*	244	245	246	232	303*
Travelguides	268*	434*	448*	307*	314*	385	489*
FB	1,264	1,501*	1,205	1,327	1,333	1,373	1,494*

while visually, concreteness shows two noteable "peaks" (Fig. 2).[42] Overall, we can confirm the difference between groups for the three features previously tested [8], while adding the observation of slight differences also for language heavy in visual and haptic information, as well as a robust difference for interoceptive information in the implicit group.

For reference, we conducted the same experiment on our reference corpora *EmoBank* and *FB*, dividing the data into implicit and explicit groups of sentences as outlined in Section 4.2. These results are reported in table 4. Histograms to support this difference in feature values between groups in *EmoBank* can be found in Appendix, Fig. 7 – as in *Fiction4*, levels of arousal, dominance and interoceptive are lower in the implicit group, while concreteness is higher in the implicit group. In the reference corpora (table 4), the strongest effect size is tendentially, as in *Fiction4*, interoceptive values. Notably, interoceptive and arousal values hold significant discriminating power across all corpora, as well as concreteness if we disregard the *FB* corpus. We see important differences within the *EmoBank*, where all features appear important only in the more personal or imaginative genres, and not in newspaper and essays.

5.3. Experiment 2

In the second experiment, we check for correlations (rather than simple statistical difference) between human/model disagreement and the level of each semantic trait per sentence. This allows us to observe whether the presence of "undetected sentiment" in text has a linear relation with any of the semantic dimensions selected. For this, *we used only sentences found in the implicit group* (as outlined in Section 4.2), so that we correlated the amount of disagreement

[42]The results of the Mann-Whitney U test are supported by a linear regression, where we sought to model the two groups by each feature. Significant results of the linear regression correspond to those indicated by the Mann-Whitney U-test, see Table 8 in Appendix.

Table 5
Spearman's ρ between disagreement (absolute human/RoBERTa score difference) and feature score in the *FB*, *EmoBank*, and *Fiction4* corpora. Note that for these correlations, we have filtered out sentences shorter than five words. For all correlations in black: $p < .05$, with *: $p < .01$.

	Concret.	Arousal	Dominance	Imageab.	Visual	Haptic	Interocept.
FB	0.03	-0.04	0.17*	-0.05	-0.02	-0.02	0.03
EmoBank	0.06*	0.01	0.05*	0.03	0.09*	0.02	-0.10*
Letters	0.01	0.02	0.06	0.02	0.02	-0.04	-0.03
Blog	0.16*	-0.12*	-0.07	0.06	0.16*	0.10	0.07
Newspaper	-0.03	0.02	0.06	0.02	0.09	0.08	0.01
Essays	-0.13	-0.01	0.16*	-0.15*	-0.05	-0.03	-0.09
Fiction	0.04	-0.08	-0.06	0.01	0.06	0.02	-0.12*
Travelguides	0.17*	0.08	0.03	0.15*	0.03	0.12*	0.01
Fiction4	-0.05	-0.09*	0.12*	0.01	0.01	-0.05	0.05
Hymns	-0.02	-0.05	0.22*	0.06	0.01	-0.03	0.01
Fairy Tales	0.05	-0.16*	-0.03	-0.02	0.06	-0.02	-0.25*
Prose	-0.07	-0.08*	0.06	0.03	-0.02	-0.02	0.04
Poetry	0.02	-0.13*	0.09	0.02	0.07	-0.04	0.04

between human and model with our chosen features in sentences where humans perceived sentiment, but models did not. We report our results in Table 5.

First of all, not all patterns of sentiment implicitness as seen in Experiment 1 are detectable as a correlation, suggesting that some of these features do not impact sentiment evocation linearly. On the other hand, we do see correlations that point to interesting genre differences in how sentiment is perceived in texts. For *Fiction4*, we find a consistent negative correlation between human/model disagreement and arousal which aligns with the lower levels of arousal in the implicit group we saw in Experiment 1. While concreteness and interoception do not show consistent linear correlations with disagreement, effects of low interoception related to higher disagreement are evident in Andersen's fairy tales. [43]

Within *Fiction4*, the role of high concreteness paired with higher dominance, and, to a lesser extent, lower arousal in sentiment disagreement (which we link to evocation) is confirmed, as well as the negative correlation of disagreement with interoception.

For comparison, we redid correlations in the reference corpora detailed in Section 3.1. Here, positive correlations are also found with concreteness: the more concrete a sentence is, the more our SA model's sentiment judgment will differ from that of human's. The strongest role of concreteness in sentiment disagreement appears to be not in literary texts proper, but in the travel guides and letters contained in *EmoBank*, and in blogs. Interoception also holds a negative correlation with disagreement Fiction category in *EmoBank*, as it did with Fairy tales in *Fiction4*.

Interestingly, the negative correlation of arousal with disagreement is not as consistent in the reference corpora, where we only see a negative correlation in Blogs. We find spurious positive correlations of disagreement with dominance, and visual – notably, however, correlations where they appear tend to have the same negative or positive direction across all corpora (including

[43]The absence of a linear relation with concreteness is particularly interesting, given the results in Experiment 1. Concreteness appears to have an effect on the evoked sentiment for human readers, but the two elements are not systematically related - the evoked sentiment does not change linearly with an increase in concreteness.

Table 6
Mean and SD of feature scores for different types. Since feature values show very slight difference across genres, to represent "highwater marks", we have added the mean values of image-captions and free emotion event descriptions (Section 3.1). Numbers in green represent the highest and in red the lowest values.

	Concret.	Arousal	Dominance	Imageab.	Interocept.	H/R disagr.
FB	2.70 ± 0.45	0.46 ± 0.11	0.52 ± 0.11	356.74 ± 58.47	1.43 ± 0.53	0.32 ± 0.20
EmoBank	2.65 ± 0.42	0.44 ± 0.10	0.54 ± 0.10	339.60 ± 49.89	1.10 ± 0.47	0.38 ± 0.20
Letters	2.68 ± 0.40	0.45 ± 0.08	0.57 ± 0.09	349.75 ± 52.19	1.13 ± 0.39	0.33 ± 0.23
Blog	2.61 ± 0.45	0.45 ± 0.10	0.54 ± 0.10	331.75 ± 52.54	1.12 ± 0.46	0.38 ± 0.20
Newspaper	2.62 ± 0.32	0.46 ± 0.08	0.57 ± 0.09	329.61 ± 40.49	0.93 ± 0.33	0.39 ± 0.20
Essays	2.49 ± 0.33	0.45 ± 0.09	0.55 ± 0.08	317.27 ± 41.25	0.96 ± 0.33	0.41 ± 0.19
Fiction	2.69 ± 0.47	0.43 ± 0.11	0.50 ± 0.11	349.44 ± 50.15	1.32 ± 0.55	0.39 ± 0.20
Travelguides	2.81 ± 0.43	0.43 ± 0.08	0.54 ± 0.07	349.25 ± 50.04	0.83 ± 0.27	0.39 ± 0.22
Fiction4	2.72 ± 0.46	0.43 ± 0.12	0.51 ± 0.13	353.90 ± 50.88	1.26 ± 0.53	0.35 ± 0.21
Hymns	2.58 ± 0.43	0.45 ± 0.13	0.56 ± 0.13	351.46 ± 47.78	1.39 ± 0.54	0.30 ± 0.21
Fairy tales	2.70 ± 0.37	0.42 ± 0.09	0.50 ± 0.11	349.99 ± 39.55	1.18 ± 0.47	0.34 ± 0.21
Prose	2.72 ± 0.36	0.42 ± 0.11	0.49 ± 0.10	348.36 ± 37.53	1.22 ± 0.45	0.39 ± 0.20
Poetry	2.90 ± 0.56	0.42 ± 0.14	0.47 ± 0.12	365.96 ± 69.00	1.16 ± 0.59	0.38 ± 0.19
Captions	3.12 ± 0.36	0.42 ± 0.10	0.51 ± 0.09	384.71 ± 52.78	0.81 ± 0.28	-
Emotion ev.	2.60 ± 0.30	0.46 ± 0.09	0.51 ± 0.09	349.85 ± 33.50	1.32 ± 0.34	-

Fiction4), with the exception of imageability. Facebook posts in *FB* seem to have no significant link to many of these channels.

5.4. Genre differences

While most datasets seem to exploit some form of trade-off between concreteness, on one side, and arousal, dominance and interoceptive on the other, relatively few show correlations with the visual and haptic semantic information, as well as with imageability. The exceptions are blogs in *EmoBank*, which shows a weak but positive correlation with visual and with human/model disagreements; and travel guides where disagreement has a correlation with haptic and imageability. *FB* and the *EmoBank* newspaper and letter category return non-significant correlations with most dimensions, with the exception of dominance for *FB*.

Genre differences in the overall use of these semantic traits can be observed in Table 6. Note that, for example, Facebook posts seem to have high values of imageability. Still, at the same time, imageability in posts displays no correlation to the absolute disagreement between model and human (Table 5). In other words, it may be that although the language of posts is highly imageable, the images are not used in a way that subtly evokes human emotion and that challenges models as much as it happens in, for example, travel guides. Similarly, literary genres (in *EmoBank* and *Fiction4*) seem to have high values for imageability and visual scores (Table 6), but these dimensions exhibit little correlation with human/model disagreement.

5.5. Relation between features

Through all our datasets, concreteness has a positive relation with disagreement – showing higher levels where models are unable to capture the sentiment that humans perceive – as

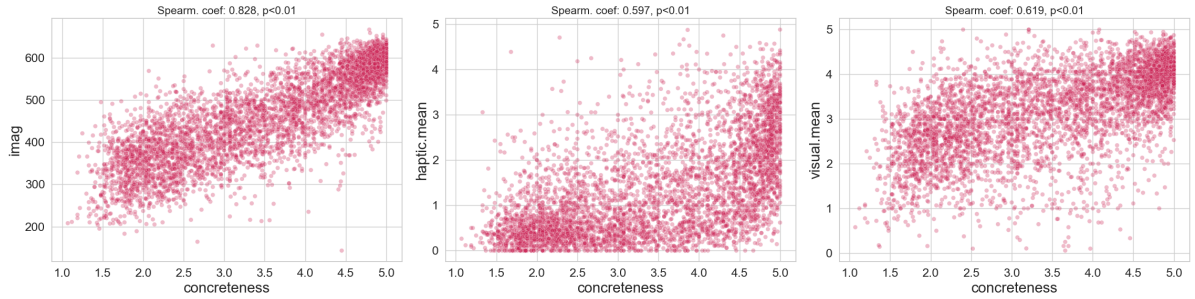


Figure 3: Correlation between concreteness and imageability, haptic and visual. Note Spearman's ρ at the top for each plot.

much as interoception has a negative one. In general, the opposite strength of concreteness and interoception in all of our datasets appears to confirm our intuition that interoception works as a sort of anti-concreteness when it comes to the evocation of sentiments, as the usage of external objects and "things" to evoke sentiments in the reader will make a low recourse to the interoceptive dimension. The fact that visual, haptic and imageablity correlations, when relevant, tend to the same direction as concreteness also adds to this hypothesis.

It is intriguing that a positive correlation between human/model disagreement and concreteness tend to co-occur with a positive correlation with sensory norms. The intuition that what is concrete as something "that is perceived through the senses" or "that can be drawn" would have led us to expect correlations of, e.g., haptic and concreteness to co-occur. On the other hand, concreteness does not have to occur with explicit sensory information at all: many words like *house*, *sea*, or *wood*, do not peak on one specific sense, and yet are considered fairly concrete; and some words like *melody* or *rhythm* might be less concrete and yet have sensory associations. The kind of concreteness that matters here might be more related to a general physical materiality rather than to a specific sensory load.

Concreteness exhibits a stronger correlation with visual, haptic, interoceptive and imageability traits ($\rho > .5,\ p < .01$) than with dominance and arousal (around $\rho = .2,\ p < .01$). When correlating terms in the concreteness dictionary with other semantic traits, we find that especially interoceptive and concreteness show an interesting correlation. Words referring to internal emotional states, also presumbaly having higher arousal (e.g., "lovesickness", "hopelessness"), tend to cluster in the direction of high interoception and low concreteness. While concreteness has a robust negative correlation with interoception ($\rho = -.52,\ p < .01$), i.e., high concreteness words generally are less interoceptive, we find that there is a set of words which maintain high interoception and high concreteness (upper right in Fig. 4). These words may be characterized as referring to concrete objects, which are nevertheless associated with internal (vs. external) states and experiences (e.g. "bladder", "breath"). Conversely, words in the lower right corner, with high concreteness and low interoceptive values, appear to be more of the category objects of external experience (e.g. "jewel", "clip", "lightswitch"), less associated with internal sensation (than is, e.g., "bee sting" or "drugs"). Considering the opposite correlation of human/model disagreement with interoception vs concreteness, we hypothesize that words used in instances of "objective correlative" would predominantly appear in the lower right corner of Fig. 4). This means that they are associated with words that are not only more concrete but also "objective" in the sense of referring to external rather than subjective or internal experiences. Therefore, the "objective correlative" might be understood

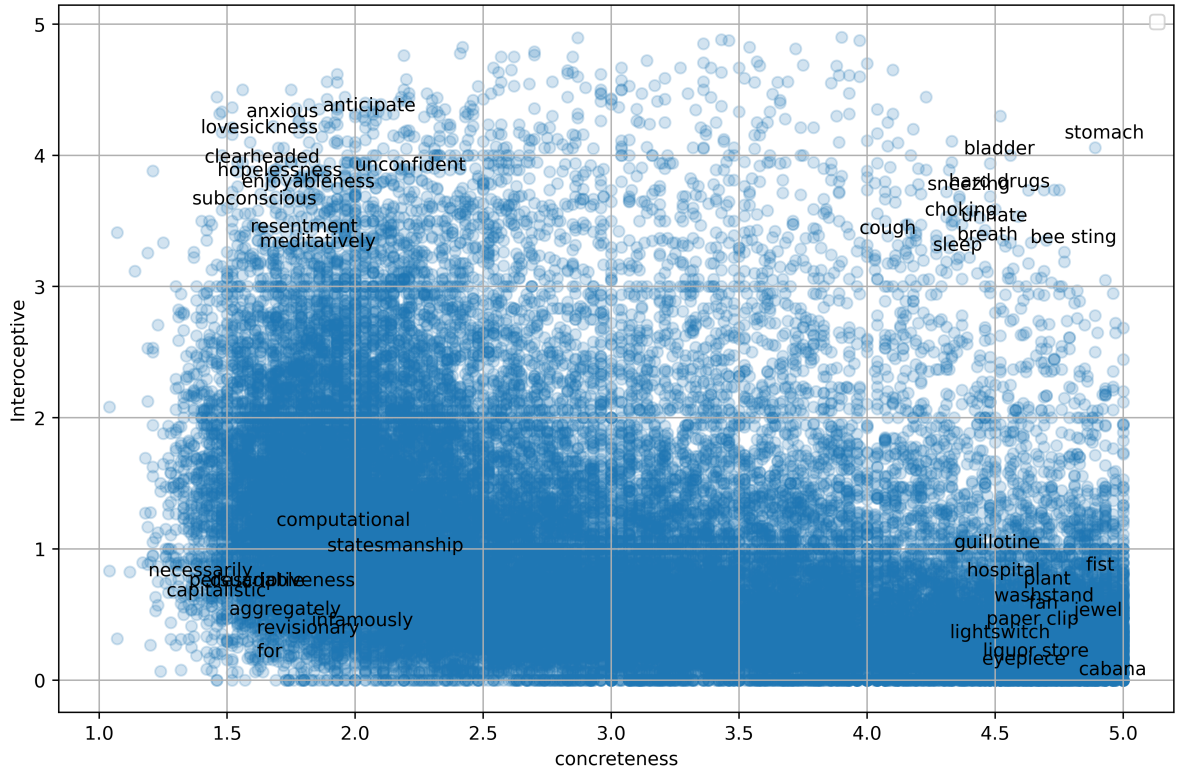


Figure 4: Overlap between the interoceptive and concreteness lexica, showing the values of words in both lexica. A random set of words is visualized in each pole.

as "objective" in both senses: impersonal and focused on external objects.

6. Discussion and conclusion

We have examined the relation between human/model disagreement on sentiment annotations and a chosen set of semantic traits for a new corpus of literary prose, comparing them with datasets representing several other domains, and we have extended the semantic traits' set used in previous literature to include also the sensory and interoceptive dimensions. Overall, we confirm previous results obtained on smaller data about relation of semantic traits to the presence of "undetected" or implicit sentiment in literary fiction, and we have observed similar trends also in non-fictional domains, with interesting differences between genres.

The "undetected" sentiments are likely to be evoked, rather than stated, and this evocation seems to pass through a trade-off between several semantic traits: an increase in concreteness and a decrease in arousal and interoception.[44] These traits seem to align with what in literary theory has been called "objective correlative": the strategy of conveying sentiment (or emotion) through the reference of external, material, or "objective" reality. This seems to happen together with the downplaying of semantic dimensions related to intensity and control, contributing to a subtler, less explicit form of emotional communication, which we

[44]Perhaps surprisingly, the concreteness' effect did not have a strong link with existing norms of sensory information, but only with interoception, with the partial exception of literary prose.

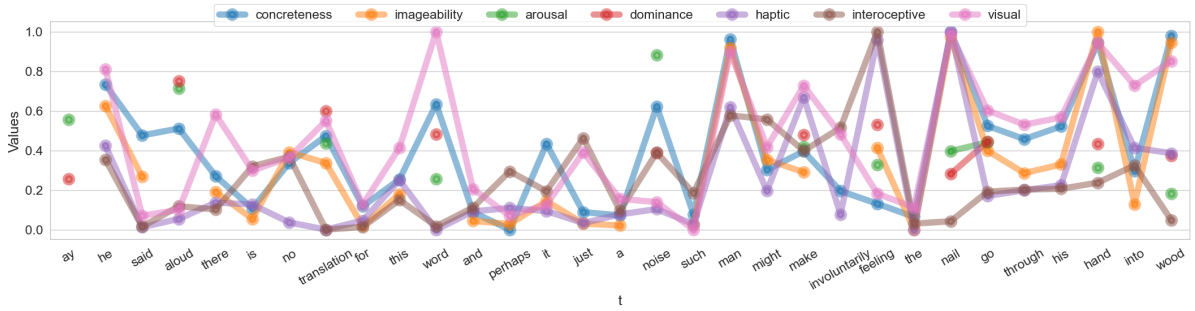


Figure 5: Standardized semantic trait values (scaled to range -1 to 1) in an example sentence. Dots connect two or more successive values (values are 'NaN' if the lemma was not in the respective feature-lexicon). The mean concreteness for this sentence = 0.4, and mean interoception = 0.25.

might characterize as an omissive evocative strategy. An example of the trade-off between omission and use of objective correlative can be observed in the following sentence of Hemingway from *Fiction4*: `` `Ay, he said aloud. There is no translation for this word and perhaps it is just a noise such as a man might make, involuntarily, feeling the nail go through his hands and into the wood'' . The sentence was consistently rated as negative by humans, and neutral by the model (see Fig. 5 above). Note how concreteness and interoception tend to divert in this sentence (e.g. on "feeling" or "nail"), while arousal and dominance values are sparse (Fig. 5).

In the future, we intend to expand our analysis to larger and more diverse corpora, and integrate more psycholinguistic resources, seeking ultimately to contribute to the development of better tools for sentiment analysis in literary genres. We would also like to observe the relation between reader response or literary reception and the concreteness-dominance or concreteness-interoception trade-off.

Limitations

We want to underline that our corpus of fiction (*Fiction4*) is limited, with only one author representing three of the four categories (Plath for Poetry, Hemingway for prose, and Andersen for fairy tales). Moreover, the demographic of our dataset is reduced (in terms of gender, ethnicity, age, social class, etc.). Replication of these results on a larger and more diverse corpus of fiction is needed, and our results should be interpreted with this in mind.

Online Resources

See https://github.com/centre-for-humanities-computing/literary_evocation for code and data.

Acknowledgments

We want to thank everyone who contributed to this work, especially Mia Jacobsen, as well as colleagues and friends for pointing out pitfalls and sharing ideas.

References

[1] S. Rebora, Sentiment analysis in literary studies. A critical survey, Digital Humanities Quarterly 17 (2023). URL: https://www.digitalhumanities.org/dhq/vol/17/2/000691/000691.html#kim-klinger2018b.

[2] E. Kim, R. Klinger, A survey on sentiment and emotion analysis for Computational Literary Studies, Zeitschrift für digitale Geisteswissenschaften (2019). URL: http://arxiv.org/abs/1808.03137. doi:10.17175/2019_008.

[3] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, P. S. Dodds, The emotional arcs of stories are dominated by six basic shapes, EPJ Data Science 5 (2016) 1–12. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0093-1. doi:10.1140/epjds/s13688-016-0093-1.

[4] M. Jockers, A Novel Method for Detecting Plot, 2014. URL: https://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/, Matthew L. Jockers Blog.

[5] K. Elkins, The Shapes of Stories: Sentiment Analysis for Narrative, Cambridge University Press, 2022. doi:10.1017/9781009270403.

[6] Y. Bizzoni, P. Feldkamp, 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, Association for Computational Linguistics, Tokyo, Japan, 2023, pp. 219–226. URL: https://rootroo.com/downloads/nlp4dh_iwclul_proceedings.pdf.

[7] S. Rebora, M. Lehmann, A. Heumann, W. Ding, G. Lauer, Comparing ChatGPT to human raters and sentiment analysis tools for german children's literature, in: A. Sela, F. Jannidis, I. Romanowska (Eds.), Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023, volume 3558 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 333–343. URL: https://ceur-ws.org/Vol-3558/paper3340.pdf.

[8] Y. Bizzoni, P. Feldkamp, Below the sea (with the sharks): Probing textual features of implicit sentiment in a literary case-study, in: V. Pyatkin, D. Fried, E. Stengel-Eskin, A. Liu, S. Pezzelle (Eds.), Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, Association for Computational Linguistics, Malta, 2024, pp. 54–61. URL: https://aclanthology.org/2024.unimplicit-1.5.

[9] E. Savinova, F. Moscoso Del Prado, Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements, in: J. Barnes, O. De Clercq, R. Klinger (Eds.), Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 305–314. doi:10.18653/v1/2023.wassa-1.27.

[10] R. Jakobson, Linguistics and poetics, in: Linguistics and Poetics, De Gruyter Mouton, 2010 (1981), pp. 18–51. doi:10.1515/9783110802122.18.

[11] L. M. Rosenblatt, The literary transaction: Evocation and response, Theory Into Practice 21 (1982) 268–277. URL: https://www.jstor.org/stable/1476352.

[12] W. C. Booth, The Rhetoric of Fiction, 2nd edition ed., University of Chicago Press, Chicago, 1983.

[13] J. Mukařovský, Standard language and poetic language, in: P. L. Garvin (Ed.), A Prague School Reader on Esthetics Literary Structure, and Style, 1932. Georgetown University Press, 1964, pp. 17–30.

[14] D. Attridge, Peculiar Language, Routledge, 1988.

[15] C. Brooks, The well wrought urn: studies in the structure of poetry, Harcourt, 1947.
[16] I. A. Richards, Principles of Literary Criticism, Routledge, 2003.
[17] V. Rentoumi, G. Giannakopoulos, V. Karkaletsis, G. A. Vouros, Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach, in: G. Angelova, R. Mitkov (Eds.), Proceedings of the International Conference RANLP-2009, Association for Computational Linguistics, Borovets, Bulgaria, 2009, pp. 370–375. URL: https://aclanthology.org/R09-1067.
[18] M. Sandri, E. Leonardelli, S. Tonelli, E. Jezek, Why don't you do it right? analysing annotators' disagreement in subjective tasks, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2428–2441. doi:10.18653/v1/2023.eacl-main.178.
[19] E. Stengel-Eskin, J. Guallar-Blasco, B. Van Durme, Human-model divergence in the handling of vagueness, in: M. Roth, R. Tsarfaty, Y. Goldberg (Eds.), Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language, Association for Computational Linguistics, Online, 2021, pp. 43–57. doi:10.18653/v1/2021.unimplicit-1.6.
[20] T. Eliot, Selected Essays by T. S. Eliot, Faber & Faber, 1948.
[21] E. Pound, A few don'ts by an Imagiste, Poetry 1 (1913) 200–206. URL: https://www.jstor.org/stable/20569730.
[22] D. Zhou, J. Wang, L. Zhang, Y. He, Implicit sentiment analysis with event-centered text representation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6884–6893. doi:10.18653/v1/2021.emnlp-main.551.
[23] Z. Li, Y. Zou, C. Zhang, Q. Zhang, Z. Wei, Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 246–256. URL: https://aclanthology.org/2021.emnlp-main.22. doi:10.18653/v1/2021.emnlp-main.22.
[24] H. J. Alantari, I. S. Currim, Y. Deng, S. Singh, An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews, International Journal of Research in Marketing 39 (2022) 1–19. doi:10.1016/j.ijresmar.2021.10.011.
[25] H. Elsahar, M. Gallé, To annotate or not? Predicting performance drop under domain shift, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2163–2173. URL: https://aclanthology.org/D19-1222. doi:10.18653/v1/D19-1222.
[26] B. Ohana, S. J. Delany, B. Tierney, A case-based approach to cross domain sentiment classification, in: B. D. Agudo, I. Watson (Eds.), Case-Based Reasoning Research and Development, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, pp. 284–296. doi:10.1007/978-3-642-32986-9_22.
[27] K. Bowers, Q. Dombrowski, Katia and the sentiment snobs, 2021. URL: https://datasittersclub.github.io/site/dsc11.html, blog: Datasitter's Club.
[28] T. Schmidt, K. Dennerlein, C. Wolff, Using deep learning for emotion analysis of 18th and 19th century German plays, Fabrikation von Erkenntnis: Experimente in den Digital Humanities - (2021). doi:10.26298/MELUSINA.8F8W-Y749-UDLF.

[29] A. Allaith, K. Degn, A. Conroy, B. Pedersen, J. Bjerring-Hansen, D. Hershcovich, Sentiment classification of historical Danish and Norwegian literary texts, in: T. Alumäe, M. Fishel (Eds.), Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 324–334. URL: https://aclanthology.org/2023.nodalida-1.34.

[30] T. Schmidt, M. Burghardt, An evaluation of lexicon-based Sentiment Analysis techniques for the plays of Gotthold Ephraim Lessing, in: B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, S. Szpakowicz (Eds.), Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Santa Fe, New Mexico, 2018, pp. 139–149. URL: https://aclanthology.org/W18-4516.

[31] T. Strychacz, "The sort of thing you should not admit": Ernest Hemingway's aesthetic of emotional restraint, in: M. Shamir, J. Travis (Eds.), Boys Don't Cry? Rethinking Narratives of Masculinity and Emotion in the U.S., Columbia University Press, 2002, pp. 141–166. doi:10.7312/sham12034-009.

[32] M. Daoshan, Z. Shuo, A discourse study of the Iceberg Principle in *A Farewell to Arms*, Studies in Literature and Language 8 (2014) 80–84.

[33] S. Ahmed, The cultural politics of emotion, Edinburgh Univ. Press, 2010.

[34] B. Brown, Thing Theory, Critical Inquiry 28 (2001) 1–22. URL: http://www.jstor.org/stable/1344258.

[35] L. Oulanne, Lived Things: Materialities of Agency, Affect, and Meaning in the Short Fiction of Djuna Barnes and Jean Rhys, Ph.D. thesis, University of Helsinki, Helsinki, 2018. URL: http://ethesis.helsinki.fi.

[36] M. Fludernik, Towards a 'natural' narratology, JLSE 25 (1996) 97–141. doi:10.1515/jlse.1996.25.2.97.

[37] J. Burroway, Writing Fiction: A Guide to Narrative Craft, Little, Brown, 1987.

[38] J. Auracher, H. Bosch, Showing with words: The influence of language concreteness on suspense, Scientific Study of Literature 6 (2016) 208–242. doi:10.1075/ssol.6.2.03aur.

[39] S. Mohammad, Obtaining reliable human ratings of Valence, Arousal, and Dominance for 20,000 English words, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 174–184. doi:10.18653/v1/P18-1017.

[40] S. Ullrich, A. Aryani, M. Kraxenberger, A. M. Jacobs, M. Conrad, On the relation between the general affective meaning and the basic sublexical, lexical, and inter-lexical features of poetic texts— a case study using 57 poems of H.M. Enzensberger, Frontiers in psychology 7 (2017) 2073.

[41] J. T. Kao, D. Jurafsky, A computational analysis of poetic style: Imagism and its influence on modern professional and amateur poetry, Linguistic Issues in Language Technology 12 (2015). URL: https://aclanthology.org/2015.lilt-12.3.

[42] M. M. Maslej, R. A. Mar, V. Kuperman, The textual features of fiction that appeal to readers: Emotion and abstractness., Psychology of Aesthetics, Creativity, and the Arts 15 (2021) 272–283. doi:10.1037/aca0000282.

[43] M. Brysbaert, A. B. Warriner, V. Kuperman, Concreteness ratings for 40 thousand generally known English word lemmas, Behavior Research Methods 46 (2014) 904–911. doi:10.3758/s13428-013-0403-5.

[44] P. J. Stone, R. F. Bales, J. Z. Namenwirth, D. M. Ogilvie, The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of

information, Behavioral Science 7 (1962) 484.
[45] C. E. Osgood, G. J. Suci, Factor analysis of meaning., Journal of experimental psychology 50 (1955) 325.
[46] S.-T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, E. Del Campo, The representation of abstract words: Why emotion matters., Journal of Experimental Psychology: General 140 (2011) 14–34. doi:10.1037/a0021446.
[47] J. Meyers (Ed.), Hemingway: The Critical Heritage, Routledge, 1982.
[48] D. Ringgaard, M. R. Thomsen (Eds.), Danish literature as world literature, Literatures as world literature, Bloomsbury Academic, New York, 2017.
[49] B. Sandstrøm, Salmen - fra kampsang til lovprisning, in: V. A. Pedersen, M. Schack, K. P. Mortensen (Eds.), Dansk Litteraturs Historie 1100-1800, Gyldendal, 2007.
[50] K. F. Baunvig, Forestillede fællesskabers virtuelle sangritualer: Forskningsprojekt vil kaste lys over den kulturelle betydning af den virtuelle fællessang under corona-tiden, Tidsskriftet SANG 1 (2020) 40–45. doi:10.7146/sang.v1i1.137029, number: 1.
[51] C. P. Heaton, Style in *The Old Man and the Sea*, Style 4 (1970) 11–27. URL: https://www.jstor.org/stable/42945039, publisher: Penn State University Press.
[52] C. Molesworth, "with your own face on": The origins and consequences of Confessional Poetry, Twentieth Century Literature 22 (1976) 163–178. doi:10.2307/440682.
[53] C. Britzolakis, *Ariel* and other poems, in: J. Gill (Ed.), The Cambridge Companion to Sylvia Plath, Cambridge University Press, 2006, pp. 107–123. doi:10.1017/CCOL0521844967.
[54] A. Doche, A. S. Ross, 'Here is my shameful confession. I don't really "get" poetry': discerning reader types in responses to Sylvia Plath's *Ariel* on Goodreads, Textual Practice 37 (2023) 976–996. doi:10.1080/0950236X.2022.2082516.
[55] C. CCLM, Danske børn og unge har stort kendskab til H.C. Andersen, 2003. URL: https://dpu.au.dk/om-dpu/nyheder/nyhed/artikel/danske-boern-og-unge-har-stort-kendskab-til-hc-andersen.
[56] T. Lundskær-Nielsen, The language of Hans Christian Andersen's Fairy Tales – Compared with earlier tales, Scandinavistica Vilnensis 1 (2014) 97–112. URL: https://www.journals.vu.lt/scandinavistica/article/view/14002. doi:10.15388/ScandinavisticaVilnensis.2014.9.8.
[57] C. O. Alm, R. Sproat, Emotional sequencing and development in Fairy Tales, in: J. Tao, T. Tan, R. W. Picard (Eds.), Affective Computing and Intelligent Interaction, Springer, Berlin, Heidelberg, 2005, pp. 668–674. doi:10.1007/11573548_86.
[58] M. A. Nielsen, Salmesprog, in: Dansk Sproghistorie Bind 4. Sprog i brug., Aarhus University Press and Society for Danish Language and Literature (DSLDK), 2020.
[59] S. Buechel, U. Hahn, EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 578–585. URL: https://aclanthology.org/E17-2092.
[60] N. Ide, C. Baker, C. Fellbaum, R. Passonneau, The manually annotated sub-corpus: A community resource for and by the people, in: J. Hajič, S. Carberry, S. Clark, J. Nivre (Eds.), Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 68–73. URL: https://aclanthology.org/P10-2013.
[61] D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, E. Shul-

man, Modelling valence and arousal in Facebook posts, in: A. Balahur, E. van der Goot, P. Vossen, A. Montoyo (Eds.), Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, San Diego, California, 2016, pp. 9–15. URL: https://aclanthology.org/W16-0404. doi:10.18653/v1/W16-0404.

[62] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565. URL: https://aclanthology.org/P18-1238. doi:10.18653/v1/P18-1238.

[63] H. G. Wallbott, K. R. Scherer, How universal and specific is emotional experience? evidence from 27 countries on five continents, Social Science Information 25 (1986) 763–795. doi:10.1177/053901886025004001.

[64] F. Barbieri, L. E. Anke, J. Camacho-Collados, XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and beyond, 2022. doi:10.48550/arXiv.2104.12250.

[65] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level Sentiment Analysis, in: R. Mooney, C. Brew, L.-F. Chien, K. Kirchhoff (Eds.), Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 347–354. URL: https://aclanthology.org/H05-1044.

[66] V. Batanović, M. Cvetanović, B. Nikolić, A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts, PLoS ONE 15 (2020). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7660500/. doi:10.1371/journal.pone.0242050.

[67] E. Borelli, D. Crepaldi, C. A. Porro, C. Cacciari, The psycholinguistic and affective structure of words conveying pain, PloS one 13 (2018) e0199658.

[68] A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 english lemmas, Behavior research methods 45 (2013) 1191–1207.

[69] D. Lynott, L. Connell, M. Brysbaert, J. Brand, J. Carney, The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words, Behavior Research Methods 52 (2020) 1271–1291. doi:10.3758/s13428-019-01316-z.

[70] J. Charbonnier, C. Wartena, Predicting word concreteness and imagery, in: S. Dobnik, S. Chatzikyriakidis, V. Demberg (Eds.), Proceedings of the 13th International Conference on Computational Semantics - Long Papers, Association for Computational Linguistics, Gothenburg, Sweden, 2019, pp. 176–187. doi:10.18653/v1/W19-0415.

[71] C. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225. doi:10.1609/icwsm.v8i1.14550.

[72] M. Wan, Q. Su, K. Ahrens, C.-R. Huang, Perceptional and actional enrichment for metaphor detection with sensorimotor norms, Natural Language Engineering (2023) 1–29. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/perceptional-and-actional-enrichment-for-metaphor-detection-with-sensorimotor-norms/0BA36E2578B2AD80CCCE00E6AF6969AB. doi:10.1017/S135132492300044X.

[73] C. Kennington, Enriching language models with visually-grounded word vectors and the Lancaster Sensorimotor Norms, in: A. Bisazza, O. Abend (Eds.), Proceedings of the 25th

Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 148–157. doi:10.18653/v1/2021.conll-1.11.

[74] M. Coltheart, The MRC Psycholinguistic Database, The Quarterly Journal of Experimental Psychology Section A 33 (1981) 497–505. doi:10.1080/14640748108400805.

[75] A. Gargett, J. Ruppenhofer, J. Barnden, Dimensions of metaphorical meaning, in: M. Zock, R. Rapp, C.-R. Huang (Eds.), Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 166–173. doi:10.3115/v1/W14-4721.

Table 7
Overview of texts within the four genres of the *Fiction4* corpus: the total number of lines – which are verses (in the case of hymns and poetry) and sentences (in the case of fairy tales and prose) – the number of words per dataset, the mean number of lines (verses/sentences) per text, the year of publication, and the number of human annotators. Note that the number of lines also represents the number of annotations, as they were done on a verse/sentence basis.

	Texts	**Lines**	**Words**	**$\bar{x}$ Words/Line**	**Period**	**Annotators**
Hymns	65	2,026	12,798	6.3	1798-1873	2
Fairy tales	3	771	18,597	2.1	1837-1847	3
Prose	1	1,900	30,279	15.7	1952	2
Poetry	40	1,545	11,576	7.3	1965	3
Full *Fiction4*	109	6,300	73,250	11.6	1837-1965	>2

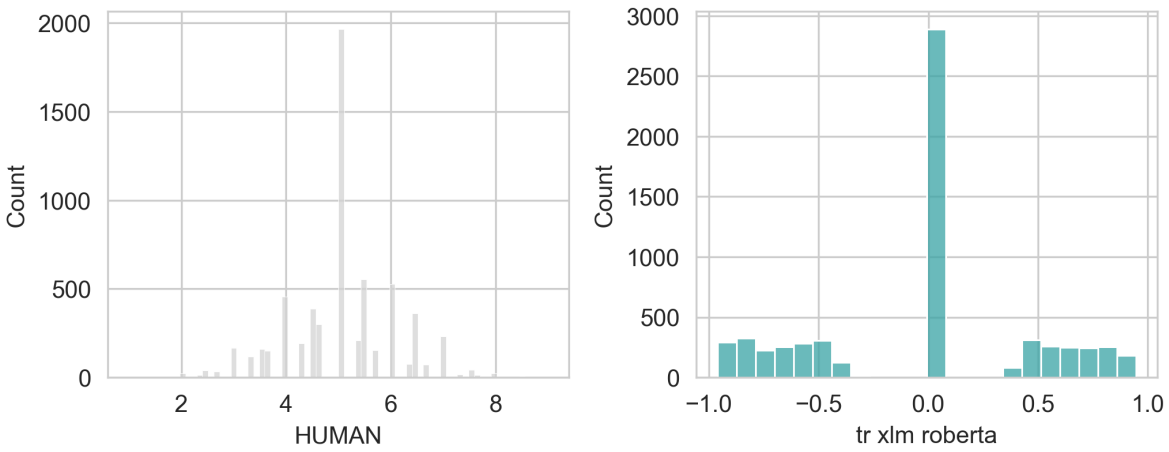
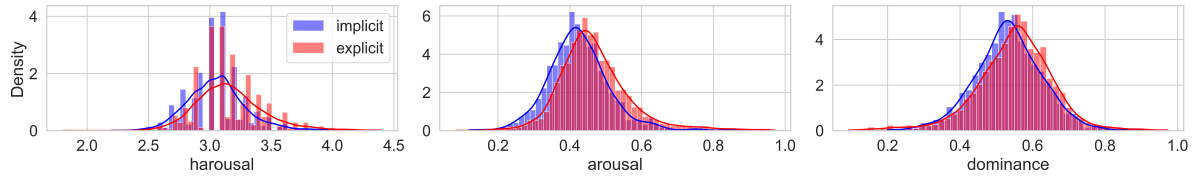


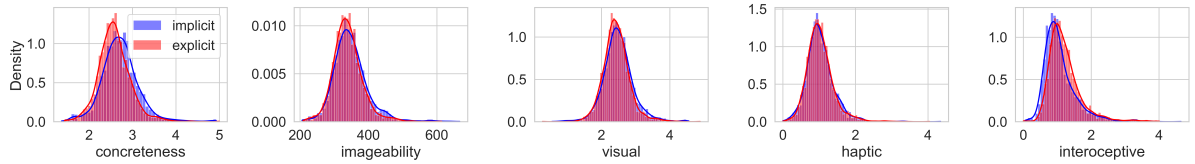
Figure 6: Distribution of human and RoBERTa scores for the *Fiction4* corpus.

Table 8
Effect sizes of the Mann-Whitney U test on the implicit vs. explicit group of sentences for each genre of the *Fiction4* corpus and the R^2 of the linear regression (seeking to model the two groups) for reference. While R^2 coefficients are low, the idea is to show that a difference exists between the groups in terms of feature values – which nevertheless show large overlaps visually – not to model valence as such via these features. We also include the test between groups within each genre in *Fiction4* separately, and the test for the *FB* corpus for comparison. For readability, we have divided the effect size by 100 here. Note that group sizes of the implicit/explicit sentences differ across the genre, with the smallest groups compared being fairy tales ($i = 147/e = 317$). Values in black: $p < .05$, with*: $p < .01$.

Measure	Corpus	**Concret.**	**Arousal**	**Dominance**	**Imag.**	**Visual**	**Haptic**	**Interocept.**
MW	*Fiction4*	1,278*	1,807*	1,636*	lightgrey1,538	1,413*	1,461*	1,975*
MW	Hymns	151*	228*	221*	200	163*	183	254*
MW	Fairy tales	20*	24	23	21	22	22	26*
MW	Prose	121*	170*	138	139	129*	133*	174*
MW	Poetry	60*	73*	59	63	68	67	86*
R^2	*Fiction4*	0.02*	0.03*	0.01*	0	0.01*	0.01*	0.05*



(a) Difference between implicit/explicit groups in arousal, dominance and human annotated arousal. We add the latter for reference since it is available in the *EmoBank* corpus. Note that harousal and arousal behave similarly.



(b) Difference between implicit/explicit groups in concreteness, imageability, visual, hapric and interoceptive levels.

Figure 7: Feature levels in the implicit and explicit groups of the *EmoBank* corpus.