In the Context of Narrative, we Never Properly Defined the Concept of Valence

Peter Boot[1,*], Angel Daza[2], Carsten Schnober[2] and Willem van Hage[2]

[1]Huygens Institute for the History and Culture of the Netherlands (KNAW), The Netherlands
[2]Netherlands eScience Center, The Netherlands

Abstract

Valence is a concept that is increasingly being used in the computational study of narrative texts. We discuss the history of the concept and show that the word has been interpreted in various ways. Then we look at a number of Dutch tools for measuring valence. We use them on sample fragments from a large collection of narrative texts and find only moderate correlations between the valences as established by the various tools. We discuss these differences and how to handle them. We argue that the root cause of the problem is that Computational Literary Studies never properly defined the concept of valence in a narrative context.

Keywords

valence, polarity, sentiment, word-embedding, narrative, computational literary studies

1. Introduction

The study of emotion and sentiment is increasingly popular in computational literary studies [20]. In this paper we will look specifically at the concept of valence, the positive or negative sentiment associated with a word or a text passage. The concept is used in a number of recent studies, but it seems to be used in quite different ways. This calls for a deeper look at the history of the concept.

One of the most-quoted studies in the field is the article by Reagan et al. about six basic shapes in the emotional arcs in stories [30]. The emotional arcs that the authors create represent the flow of what they call 'sentiment', measured by the 'Hedonometer', a dictionary-based tool that assigns sentiment to words [11]. There is no discussion in the article about the status of this sentiment: does it correspond to the sentiment that readers experience? Does it correspond to sentiment of the characters or the narrator? The paper includes an annotated emotion arc for *Harry Potter and the Deathly Hallows* which seems to show that the highs and lows of the story correspond to the maxima and minima in the arc. But the paper treats the construction of the arcs on the basis of the sentiment data as a purely technical problem, without asking what it is exactly that these arcs are modelling.

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

[*]Corresponding author.
peter.boot@huygens.knaw.nl (P. Boot); j.daza@esciencecenter.nl (A. Daza); c.schnober@esciencecenter.nl (C. Schnober); w.vanhage@esciencecenter.nl (W. v. Hage)
0000-0002-7399-3539 (P. Boot); 0000-0003-1711-3151 (A. Daza); 0000-0001-9139-1577 (C. Schnober); 0000-0002-6478-3003 (W. v. Hage)

Bizzoni and Feldkamp [2] do address the issue in a case study on *The Old Man and the Sea*. For each sentence in the novel, two annotators rated 'the sentiment expressed by the sentence'. They were instructed 'to avoid rating how a sentence made them feel and to try to report only on the sentiments actually embedded in the sentence, i.e., to think about the valence of each sentence individually, without overthinking the story's narrative to reduce contextual interpretation'. This is an interesting instruction, in that it explicitly states the sentiment is not in the reader and it shouldn't relate to the story events. It assumes that there is such a thing as '*the* sentiment embedded in a sentence'. The paper then goes on to check whether LLM or dictionary-based sentiment models correlate with the annotators' ratings. We will come back to this study below.

Rebora, in his survey of sentiment analysis in literary studies [31], also asks where the sentiment is supposed to reside: in the text or in the reader. He notes that in literary studies, narratologists would opt for the text, while students of reader response would look at the reader. But he also points at a third possibility: the characters as vehicles of emotion. Nalisnick and Baird, e.g., have applied sentiment analysis to Shakespeare's plays in order to study the relations between the plays characters [27]. And there are other possibilities: the sentiment that one finds in the text can also be used to gauge the sentiment of the author, as in Stirman and Pennebaker's study of suicidal poets [35]. It is even possible to study the sentiment in novels and other texts not out of an interest in anything that has to with the book itself, but to read larger social attitudes. For example, in their study of the perception of coal and oil in recent U.S. works [14], Grubert and Algee-Hewitt are interested in the perception of these energy sources in contemporary U.S. society.

It is clear that all of the above approaches to sentiment can be enlightening. But there is a danger that we forget that sentiment in texts can have these different aspects, and our current tooling is certainly not able to distinguish them. In an effort to create some clarity, we will in this paper briefly recount the history of the concept of valence, focusing on the various definitions researchers have used as well as on how it was established or computed. In a second, empirical part, we look at a number of tools for computing valence in Dutch. In so far as they are dictionary-based we look at the overlap between dictionaries and the degree to which they assign the same values to their shared words. Then we use a sample of fragments from Dutch fiction to assess how the various tools compare in their assignment of valence to these fragments.[1]

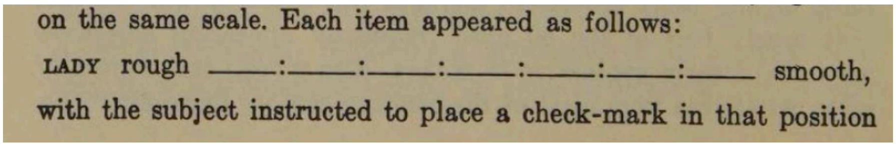
Note that our main interest here is not in the sentiment arcs that can be derived from book segments' valences (see also [12]). What we want to contribute to is the much more elementary question: what is a word or chunk valence in the first place, and how do we compute it?

2. Background: Valence, Sentiment and Polarity

2.1. Valence in psycho-linguistic studies

We start our short history of the concept of valence with the 1957 book *The measurement of meaning*, by Osgood, Suci and Tannenbaum [29]. Osgood and his colleagues wanted to describe

[1] The notebooks and datasets underlying this paper are available at https://doi.org/10.5281/zenodo.13942218.



on the same scale. Each item appeared as follows:

LADY rough ____:____:____:____:____:____:____ smooth,

with the subject instructed to place a check-mark in that position

Figure 1: Lay-out of Osgood et al.'s questions. *The Measurement of Meaning*, p. 34. Image from the Internet Archive.

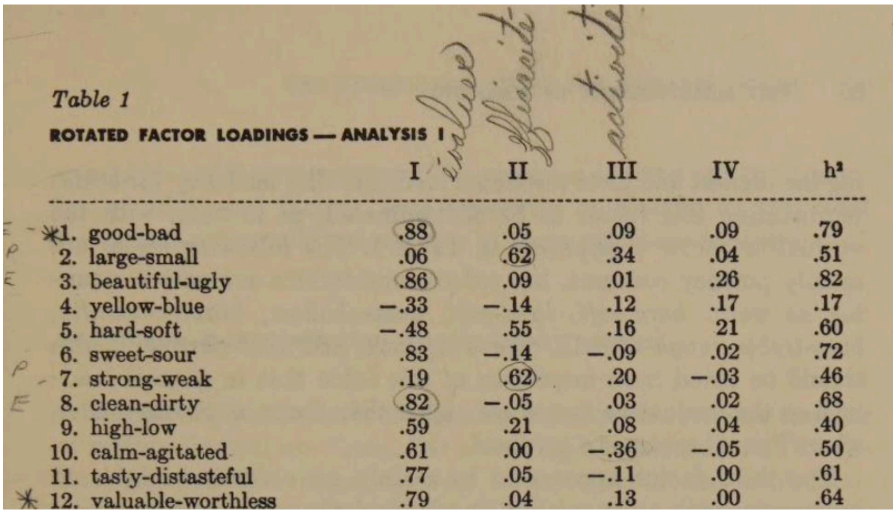


Table 1

ROTATED FACTOR LOADINGS — ANALYSIS I

évaluative puissance activité

	I	II	III	IV	h²
E - ✱1. good-bad	.88	.05	.09	.09	.79
P 2. large-small	.06	.62	.34	.04	.51
E 3. beautiful-ugly	.86	.09	.01	.26	.82
4. yellow-blue	−.33	−.14	.12	.17	.17
5. hard-soft	−.48	.55	.16	21	.60
6. sweet-sour	.83	−.14	−.09	.02	.72
P - 7. strong-weak	.19	.62	.20	−.03	.46
E 8. clean-dirty	.82	−.05	.03	.02	.68
9. high-low	.59	.21	.08	.04	.40
10. calm-agitated	.61	.00	−.36	−.05	.50
11. tasty-distasteful	.77	.05	−.11	.00	.61
✱ 12. valuable-worthless	.79	.04	.13	.00	.64

Figure 2: Factor loadings in Osgood et al. *The Measurement of Meaning*, p. 37. Image with student notes from the Internet Archive.

the meaning of certain concepts, such as the word 'lady', and asked test subjects to associate these nouns with positions on a scale between opposite adjectives, such as good vs. bad, hard vs. soft or kind vs. cruel. They used a Likert scale, and the form might look as in Figure 1.

After averaging, this gave them, for twenty concepts and fifty pairs of adjectives, 1000 measurements. A factor analysis identified the hidden dimensions underlying the measurements. A fragment of the resulting table is reproduced in Figure 2. We see that the results of some adjective pairs, such as good vs. bad and beautiful vs. ugly, are almost completely explained by the hidden variable I, the scores on strong vs. weak are mostly explained by hidden variable II, etc. But what are these hidden variables? Osgood et al. then write 'The problem of labelling factors is somewhat simpler here than in the usual case. (...) The first factor is clearly identifiable as *evaluative* (...). The second variable identifies itself fairly well as a *potency* variable (...). The third factor appears to be mainly an *activity* variable (...)' (italics original).

Later, these dimension would become known by other names: evaluativeness as pleasure or valence, activity as arousal, potency as power or dominance. This is not the whole story, but Osgood and colleagues made a fundamental contribution to the three-factor theory of emotion. It is interesting to note that they already mention that these factor loadings depend on cultures: e.g., for Japanese and Korean respondents, the adjective pair delicate vs. rugged clearly belonged to the evaluative dimension, for U.S. respondents it did not.

We continue with a look at *The General Inquirer*, one of the first text analysis programs [36].

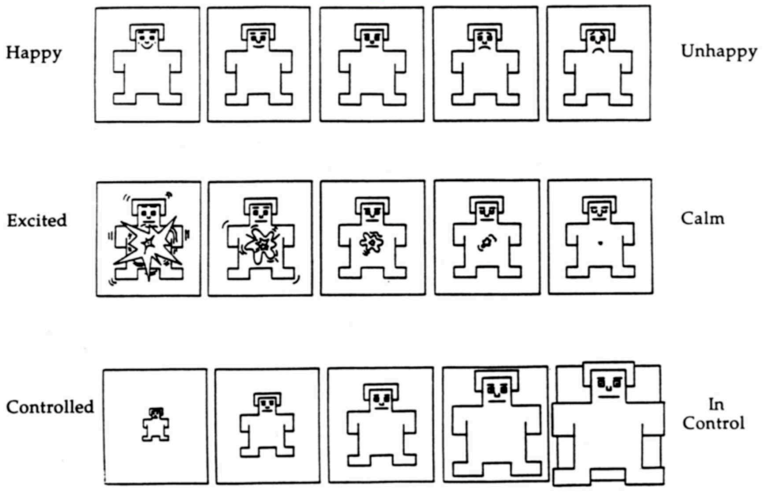


Figure 3: Self-assessment mannikin.

The 1966 book describes among others the Harvard III dictionary, developed for use with the General Inquirer. Four groups of words were included to account for the high and low ends of the evaluativeness and potency variables found by Osgood (pp. 176, 185). We see here that what for Osgood were hidden variables, the result of a computational process that required the interpretative step of labelling, have now become measurable entities underlying texts.

A next step in our tale was set by Bradley and Lang in 1999, in their paper 'Affective Norms for English Words' [7]. As prompts in psychological research they needed words with known values for the dimensions of (what they called) Pleasure, Arousal and Dominance. They asked subjects to rate the words using a 'self-assessment mannikin' (see Figure 3 for an example). For pleasure, the instruction that they gave their subjects was 'At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful. *When you feel completely happy* you should indicate this by bubbling in the figure at the left. The other end of the scale is when you feel completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored' (italics ours). The valence associated with a word now equals (possibly) complete happiness or unhappiness of the subject, i.e., a subjective feeling.

The creation of ever larger dictionaries of words with associated valence and other variables has been a constant in psycholinguistic research ever since. We mention a few: Warriner and colleagues created a list of almost 14000 English lemmas with Valence, Arousal and Dominance [40]. They used Likert scales rather than the self-assessment mannikins, but they stuck to the subjective language: 'At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful. When you feel completely happy you should indicate this by choosing rating 1'. One reason why their work is remarkable is that they found that there are systematic differences between men and women in how they rate various categories of words. The last study for English words that we mention here is that by Mohammad [24]. Mohammad used a procedure called best-worst scaling where respondents were asked: 'Which of the four words below is associated with the MOST happiness / pleasure / positiveness / satisfaction / contentedness /

hopefulness OR LEAST unhappiness / annoyance / negativeness / dissatisfaction / melancholy / despair?' The choice of words 'is associated with' is less subjective and seems to ask for a more objective relation between the words and the associated values than formulations such as 'when you feel completely happy'. Still, Mohammad too found important differences between various groups of people (by gender, age and self-assessed Big 5 personality characteristics) with respect to the values they associate with the various words.

One study, especially relevant for the second part of this paper, was done by Moors and colleagues [26]. They created a list of 4300 Dutch words with associated valence and other values. Participants 'were asked to judge the extent to which the words in the study referred to something that is positive/pleasant ("positief/aangenaam") or negative/unpleasant ("negatief/onaangenaam")'. Introducing a third perspective, rather than giving a subjective response to a word (Bradley) or a judgment about the language system (Mohammad), subjects were now asked about properties of the objects in the world that the words represent. Moors presents separate valence scales for the general population, for women and for men. The examples that we could give of differences in evaluation between women and men are all distressingly predictable.

2.2. Valence or sentiment in consumer reviews

With the appearance of consumer reviews, the problem of establishing the opinions that they expressed and the related sentiments became important subjects for marketeers [21]. As an example of early studies we mention Hu and Liu's work [15]. Hu and Liu find sentences in reviews that contain both a product feature and an opinion. The opinion words and their polarity (= whether they express a positive or negative sentiment) are deduced from WordNet by starting with some seedwords. Over the years, their work has produced a 6800-word opinion lexicon. In the study of consumer reviews, the question is no longer how a word strikes the reader, but what was the opinion that a writer wanted to express. That also means that the attention moves to adjectives rather than the nouns that the tradition established by Bradley [7] was typically interested in.

A Dutch sentiment dictionary with a focus on review sentiment was created by De Smedt and Daelemans [8]. They used frequently occurring adjectives from book reviews, and asked annotators 'to classify each adjective in terms of positive-negative polarity and subjectivity'. The question here no longer refers to subjective feeling or properties of an object, but to a property of the sentiment word. The dictionary is special (among sentiment dictionaries) in that it distinguishes between multiple word senses. E.g. the word 'scherp' (sharp) applied to a sound has negative polarity, but in 'a sharp thinker' the word's polarity is positive. Using computational tools and based on several linguistic resources, the initial list of words has been expanded to 5500 word senses. De Smedt and Daelemans' dictionary can be used as part of the Pattern toolset [9].

2.3. Valence from word embeddings

Under the name of SentiArt, Arthur Jacobs introduced a completely different way of estimating valence in the field of computational literary studies [17]. The method uses a set of positive and negative seed words in combination with a word embedding. The valence of a word is then computed summing its similarities to the positive seed words and subtracting the similarities to the negative ones. The method is based on the assumption that, in the word embedding, words with similar meanings cluster together. Because there is no need for human ratings, this seems like an objective procedure. However, the choice of the texts that are used to create the word embedding, the procedure that is used for its computation as well as the choice of seed words are to some extent arbitrary. It is also not self-evident that the procedure should work at all. Jacobs [17, p. 3] states that the method was able to explain 34% of the variance between words found by Warriner [40], which isn't exactly promising.

2.4. Valence beyond the word

The various approaches to the valence concept that we have discussed all assign valence at the word level. It is not self evident that it is possible to define valence at a higher level, e.g. that of sentences, paragraphs or even book chapters.

Bradley and Lang extended their work on word valence to small texts (one to a few sentences) [6] These small texts describe, in the second person, situations that would probably cause an emotional state with a certain valence (as well as arousal and dominion). We give an example with low valence: 'You gag, seeing a roach moving slowly over the surface of the pizza. You knock the pie on the floor. Warm cheese spatters on your shoes'.

Specifically in the context of narrative, Rebora asked students to rate the sentiment in paragraphs of a story by Pirandello [32]. Their agreement was very weak. As we saw, in [2] two researchers rated sentiment in sentences in *The Old Man and the Sea*. Their result was better than Rebora's: the correlation between their ratings was strong, after detrending even very strong. Kaakinen and colleagues report on a multilingual database of short stories (ca. 1,000 characters) [19], for which raters established valence and arousal using self-assessment mannikins. Finally, Jacobs [16, pp. 117-119] briefly reports on a number of experiments where readers were asked to rate the valence of sentiments or short sections of among others a Harry Potter novel and *Pippi Longstocking*. He does not report agreement measures, but the resulting sentiment arcs could be predicted reasonably well using his SentiArt toolset.

Outside of the domains of psychology or narrative, there exists a plethora of studies on the polarity of especially reviews and social media texts. We just mention work on the orientation of tweets [34], targeted not so much at establishing a text's valence, but at classifying a text as positive, negative or neutral, and work on product reviews, where beyond establishing a text's overall valence the aim is to find the aspects of a product that reviewers are positive or negative about [41].

2.5. Provisional conclusion

We have seen how the concept of valence morphed from a hidden dimension of meaning into something that could be measured in text using just a few adjectives, and further into a subjective feeling, a property of the language or a property of the things that we talk about. It is clear that in practice, these are not unrelated. If I consider peace a good thing, I'll probably feel good about the word and I'll use positive words in talking about it. But conceptually they are distinct, and the extent to which they agree in practice is an empirical question. Another thing that we learned in this short review is that valence judgments vary by gender, age, personality and culture. Anyone who confidently writes about '*the* sentiment' of a narrative work should be aware of that.

3. Method

After this brief look at the history and the operationalisation of the notion of valence, we turn to a comparison of various tools that assign valence to Dutch words and texts. Except for the approaches already discussed, we also include two transformer-based language models. These tools by themselves have no notion of valence, but have been trained to fulfill classification tasks that assign short texts to evaluative categories.

Our interest here is not in finding the tool that best approximates the 'true' sentiment value, if such a thing exists, or the gold labels, which we don't have. What we are interested in is whether these tools agree or disagree and what that says about the current state of sentiment analysis for narrative, at least for Dutch.

The tools that we use are:

LiLaH the sentiment mapping of LiLaH [22], a manually corrected version of an automatic translation of the NRC emotion lexicon [25], see A.2.2.

LIWC a general-purpose tool for text analysis [5] based on an underlying dictionary, see A.2.3.

Moors discussed above, see 2.1.

Pattern discussed above, see 2.2.

SentiArt discussed above, see 2.3.

VanRoy a transformer-based model trained on book reviews, see A.2.5.

xlm a transformer-based multilingual model [1] based on Twitter, see A.2.5.

For a further description of the tools that we will use, as well as for the settings that we apply in computing the SentiArt valences, we refer to the Appendix (A.1).

For the dictionary-based tools (that includes the tools with curated dictionaries as well as SentiArt), we first compare the dictionaries. We report a number of measures:

- dictionary size;
- number of shared words;
- word-level agreement of assigned values;
- words where the dictionaries disagree.

After the dictionary comparison, we compare the result of the tools on a sample of Dutch novels. We create the sample as follows: from a collection of 10,921 recently published Dutch books we remove non-fiction and books with less than 5000 words, then select every fifth book. For every selected book (n=2087) we select a random newline character, which usually corresponds with a paragraph start. Starting from that location we select the first hundred tokens.

Then, for each of the tools mentioned in the Appendix we compute the valence assigned to the fragment. For Moors, LiLaH and SentiArt, the valence of the text fragment is the average lemma valence for all words whose lemma occurs in the relevant lexicon. Words whose lemma does not occur in the resource are ignored. For the other tools, see the Appendix for the computational procedure.

We then compute correlations between the tools' valence assignments. For some pairs of tools we also look at fragments where the two tools produce very different results.

4. Results

In describing correlations, we use the labels proposed by Evans [13]: 0.00 - 0.19: *very weak*, 0.20 - 0.39: *weak*, 0.40 - 0.59: *moderate*, 0.60 - 0.79: *strong*, 0.80 - 1.00: *very strong*.

4.1. Comparison at dictionary level

For simplicity's sake, in this section we ignore LIWC15, as it did much worse than LIWC07 in the analysis of the fragment valences.

4.1.1. Overlap at dictionary level

Table A.4 in the Appendix presents the tool dictionary sizes and their overlap. We notice a few things: (i) The SentiArt dictionary obviously dwarfs the curated dictionaries. Because of the better coverage of the text, we might expect the SentiArt dictionary to perform better than the other ones. More importantly: (ii) the overlap between the other dictionaries is in most cases quite small, also in comparison with the number of words that they do not share. Only LiLaH and Moors share more than 1000 words (1481), and for Moors, that is about one third of the words it contains. In all other comparisons, the number of overlapping words as a fraction of a dictionary's total words is (much) smaller. It would be surprising if tools that share so small a part of their vocabularies would report similar valences.

4.1.2. Agreement at dictionary level

We look first at the agreement between the continuously valued (SentiArt, Moors and Pattern) and the binary-valued tools (LIWC07 and LiLaH), given in table 1. These look as one would expect. In all cases there is a clear distinction between the mean values of the positive and the negative words in LiLaH and LIWC07. For LIWC07, the difference between the means is somewhat larger than for LILaH. LIWC07 seems to agree better with the continuous tools than LiLaH does.

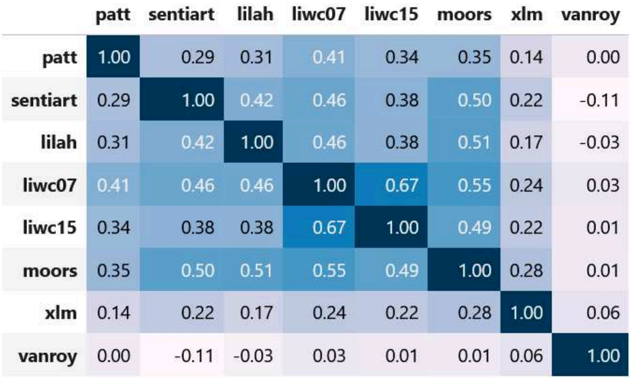
Then we look at the correlations between the continuously valued valences (Table 2). The correlations of Moors and Pattern with SentiArt are strong, the correlation of the two curated dictionaries Moors and Pattern is very strong. We should be aware, however, that these are correlations over a relatively small number of words.

Table 1
Means for the continuously valued tools' values for the positive and negative words in the binary-valued tools.

tool	stdev	LILaH pos	LILaH neg	LIWC07 pos	LIWC07 neg
Moors	1.06	4.95	2.98	5.50	2.60
Pattern	0.42	0.33	-0.34	0.52	-0.44
Sentiart	0.08	0.04	-0.07	0.07	-0.07

Table 2
Correlation for continuously valued valences.

		word overlap	correlation
Moors	Pattern	726	0.86
Moors	SentiArt	4279	0.63
Pattern	Sentiart	3223	0.69



	patt	sentiart	lilah	liwc07	liwc15	moors	xlm	vanroy
patt	1.00	0.29	0.31	0.41	0.34	0.35	0.14	0.00
sentiart	0.29	1.00	0.42	0.46	0.38	0.50	0.22	-0.11
lilah	0.31	0.42	1.00	0.46	0.38	0.51	0.17	-0.03
liwc07	0.41	0.46	0.46	1.00	0.67	0.55	0.24	0.03
liwc15	0.34	0.38	0.38	0.67	1.00	0.49	0.22	0.01
moors	0.35	0.50	0.51	0.55	0.49	1.00	0.28	0.01
xlm	0.14	0.22	0.17	0.24	0.22	0.28	1.00	0.06
vanroy	0.00	-0.11	-0.03	0.03	0.01	0.01	0.06	1.00

Figure 4: Spearman correlations of computed valences.

Finally, for all dictionary pairs we also checked whether there are words where the dictionaries disagree, and if so, whether there is an obvious culprit. For LIWC07 and Moors and to a lesser extent Pattern we hardly found apparent mistakes. LiLaH contains a number of clear errors, maybe due to the limited availability of the Dutch translator [22, p. 154]. In the SentiArt dictionary, there are countless misclassifications. See Table A.5 for examples. We also note here that among the top positive words in the SentiArt valences, there appears a curious group of words related to hospitality, such as dinner, hostess, sommelier, catering, service and culinary, which also raises some questions about the adequacy of the procedure.

4.2. Comparison at text fragment level

After computing the valences assigned to the fragments we computed their correlations (see Figure 4). We see that the correlations between the results of the various sentiment analysis tools that we have looked at doesn't get better than moderate (the only strong correlation is between the two LIWC flavours). As we don't know the 'true' valences, we're not in a position to say which is the best tool, but we can say something.

1. The correlation of the XLM transformer-based model is with the others is at best weak, for the Van Roy model there is no correlation. But we already knew that these tools were very different than the other tools, and it confirms (if confirmation were needed) that the type of training text is really important.
2. Of the other dictionaries, the agreement of Pattern with the rest is at best moderate but mostly weak. The reason is probably Pattern's background in consumer review analysis.
3. From the two LIWC dictionaries, LIWC15 performs noticeably less than LIWc07. This is probably due to the Dutch LIWC15 being an automatic translation of the English dictionary.
4. The remaining curated dictionaries (LiLaH, LIWC07 and Moors) and the SentiArt approach have moderate correlations with each other.

For the tools Moors, SentiArt and LIWC07 we also did an analysis of text fragments that scored high on one tool and low on another. The main causes for differences were apparent errors in the SentiArt word valence, homonymies, the word 'niet' (not) in Moors being assigned a low valence, and words not being present in the curated dictionaries. Some of those are unavoidable in the context of a dictionary approach, others could be avoided by better curation. Some of the options that we used for the SentiArt computations resulted from preliminary testing with these differently-rated fragments. See the Appendix for details.

5. Discussion

As the main results of the empirical part of this paper we see:

1. The SentiArt procedure to assign valence to large collections of words has serious limitations, even when computed on the basis of a domain-specific word embedding.
2. These limitations can to some extent be overcome by computing distances to a centroid vector rather than to the individual seed words, by only looking at the top and bottom quartile of the resulting valence distribution, and by excluding punctuation and function words (see Appendix for details). However, while leaving out punctuation and function words from the fragment valence computation helps, it is not a real solution to the problem that apparently the word embedding-based valence assignment is producing flawed results.
3. The agreement between other tools for computing valence of Dutch narrative text are never better than moderate.

With respect to the first two items in this list, this suggests that we may have to look beyond word2vec for a better answer to questions of semantic relatedness between lemmas (e.g. to contextual text representations as provided by Transformer-based language models [10]). The limited number of pretty arbitrary seed words seems another limitation of the SentiArt approach. A better way of obtaining valence ratings for many more words than can be manually curated might be machine learning with as target the Moors valences, and as features (among others) the word2vec distances to some of the top and bottom Moors words.

With respect to the last item, the question is: how bad is it that these tools only agree moderately? If we knew that one of the tools is mostly correct, it wouldn't matter, we could just stop using the others. But we suspect that this is not the case. We have seen enough limitations in all of the tools and in the dictionary approach as such that it is unlikely that any of these tools presents us with more than a rough approximation of correctness.

That might lead us to asking why we have focussed here on dictionary-based approaches. In their survey of sentiment analysis in literary studies, Kim and Klinger [20] wrote that 'much digital humanities research (especially dealing with text) uses the methods of text analysis that were in fashion in computational linguistics twenty years ago'. And in a direct comparison, current machine learning tools usually perform better in predicting human valence ratings (see e.g. [38] for a study by Van Atteveldt and colleagues in the field of politicology). So why do we still study these methods, rather than follow the lead of computational linguistics?

One answer to that question could come from Teodorescu and Mohammad [37], who show that, in spite of instance-level inaccuracy, dictionary-based methods work very well for larger bins of texts (e.g. for groups of 30 or a 100 tweets). They argue that '[f]or applications where simple, interpretable, low-cost, and low-carbon-footprint systems are desired, the lexicon-based systems [...] are often more suitable'. That suggests that dictionary-based methods might be better suited to study sentiment at the chapter than at the sentence level of a novel. Another answer comes from Öhman [28], who argues that for the large texts that in the humanities we are often interested in, the annotation efforts on which machine learning tools depend are are just not feasible. But the best answer is maybe one that Öhman also hints at when she proposes to leave the term 'sentiment analysis' to the computational linguists and argues that even if it is not computational sentiment, differences between texts that are made visible by dictionary-based tools, if statistically significant, are still relevant research findings.

We wouldn't go so far as as to say that what dictionary-based methods can do is not sentiment analysis. But it is true that most of the work in computational linguistics has been on the detection of sentiment in the sense of stance, where the aim is to detect the view that a text's author expresses about some object. In narrative, and especially literary narrative, the aim of the text is not to convey the author's view about the characters or the events, and if it were, it wouldn't necessarily be the aim of researchers to uncover that view. This doesn't mean that we don't need to work on well-annotated corpora of narrative on which we can apply the tools of machine learning, far from that, but it does mean that current pre-trained sentiment analysis tools have been trained on corpora so different from the corpora that we are interested in that they may not be very relevant to the analysis of narrative.

Returning to the question of how much of a problem we have with these moderate correlations, and assuming, for the sake of argument, that the correlation of our tools with the 'true' valence is about equal to the best of their mutual correlations, that is .51, what is it can we do with a measurement that misses so much information? Maybe we could look at some patterns, very carefully. But it certainly would not make sense to use these measurements as ingredients in e.g. predictive modelling or the construction of narrative arcs.

We see some ways of moving forward:

1. creating a much larger dictionary than the present curated dictionaries for Dutch along the lines sketched earlier in this section;
2. using some sort of ensemble measure, in the hope that the tools can compensate for each other's weaknesses;
3. only using sentiment analysis in a narrative context on larger text segments;
4. starting an annotation effort for valence in narrative fragments.

All of these, however, are only stop-gap measures for what we believe is the real problem, which is that we have as a discipline not really defined the concept of valence in a narrative context. As we saw in our overview of the history of the concept of word valence in section 2, many completely different definitions and operationalisations have have been proposed. It has been possible to get away with these differences in the analysis of by and large simple and straightforward texts such as social media posts. In his survey of sentiment analysis in literary studies [31], Rebora writes 'S[entiment] A[nalysis], in fact, can be performed by selecting or combining an ample variety of approaches [...]. Choosing one approach over the other means also defining the very nature of the object under examination'. We might add that to 'define the very nature of the object under examination' is what, with respect to valence, we have up to now, and to our peril, shied away from.

Acknowledgments

This research was funded by the Netherlands eScience Center, grant number ASDI.2020.032.

References

[1] F. Barbieri, L. E. Anke, and J. Camacho-Collados. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference.* Marseille, France, 2022, pp. 258–266.

[2] Y. Bizzoni and P. Feldkamp. "Comparing Transformer and Dictionary-Based Sentiment Models for Literary Texts: Hemingway as a Case-Study". In: *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages.* Tokyo, Japan, 2023, pp. 219–228.

[3] P. Boot. *LIWCTools.* Version 1.3.3. 2016. URL: https://github.com/pboot/LIWCtools.

[4] P. Boot, H. Zijlstra, and R. Geenen. "The Dutch Translation of the Linguistic Inquiry and Word Count (LIWC) 2007 Dictionary". In: *Dutch Journal of Applied Linguistics* 6.1 (2017), pp. 65–76. DOI: 10.1075/dujal.6.1.04boo.

[5] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker. "The Development and Psychometric Properties of LIWC-22". In: *Austin, TX: University of Texas at Austin* 10 (2022).

[6] M. M. Bradley and P. J. Lang. "Affective Norms for English Text (ANET): Affective ratings of text and instruction manual". In: *Techical Report. D-1, University of Florida, Gainesville, FL* (2007).

[7] M. M. Bradley and P. J. Lang. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings.* Tech. rep. Technical report C-1, the center for research in psychophysiology, 1999.

[8] T. De Smedt and W. Daelemans. ""Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives." In: *Lrec*. Istanbul, Turkey, 2012, pp. 3568–3572.

[9] T. De Smedt and W. Daelemans. "Pattern for Python". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2063–2067.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota, 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.

[11] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, et al. "Human Language Reveals a Universal Positivity Bias". In: *Proceedings of the national academy of sciences* 112.8 (2015), pp. 2389–2394. DOI: 10.1073/pnas.1411678112.

[12] K. Elkins. *The shapes of Stories: Sentiment Analysis for Narrative.* Cambridge University Press, 2022. DOI: 10.1017/9781009270403.

[13] J. D. Evans. *Straightforward Statistics for the Behavioral Sciences.* Thomson Brooks/Cole Publishing Co, 1996.

[14] E. Grubert and M. Algee-Hewitt. "Villainous or Valiant? Depictions of Oil and Coal in American Fiction and Nonfiction Narratives". In: *Energy research & social science* 31 (2017), pp. 100–110. DOI: 10.1016/j.erss.2017.05.030.

[15] M. Hu and B. Liu. "Mining and Summarizing Customer Reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* 2004, pp. 168–177. DOI: 10.1145/1014052.1014073.

[16] A. Jacobs. *Neurocomputational Poetics: How the Brain Processes Verbal Art.* Anthem Press, 2023.

[17] A. M. Jacobs. "Sentiment Analysis for Words and Fiction Characters from the Perspective of Computational (Neuro-) Poetics". In: *Frontiers in Robotics and AI* 6 (2019), p. 53. DOI: 10.3389/frobt.2019.00053.

[18] A. M. Jacobs and A. Kinder. "Computing the Affective-Aesthetic Potential of Literary Texts". In: *Ai* 1.1 (2019), pp. 11–27. DOI: 10.3390/ai1010002.

[19] J. K. Kaakinen, E. Werlen, Y. Kammerer, C. Acartürk, X. Aparicio, T. Baccino, U. Ballenghein, P. Bergamin, N. Castells, A. Costa, et al. "IDEST: International database of emotional short texts". In: *PLOS one* 17.10 (2022), e0274480. DOI: 10.1371/journal.pone.0274480.

[20] E. Kim and R. Klinger. "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies". In: *Zeitschrift für digitale Geisteswissenschaften* (2019). DOI: 10.17175/2019_008_v2.

[21] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan Claypool, 2012. DOI: 10.1007/978-3-031-02145-9.

[22] N. Ljubešić, I. Markov, D. Fišer, and W. Daelemans. “The LiLaH emotion lexicon of Croatian, Dutch and Slovene”. In: *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. Barcelona, Spain (Online), ACL, pp. 153–157, December, 2020.* Barcelona, Spain (Online), 2020, pp. 1–5.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2.* Nips'13. Red Hook, NY, USA: Curran Associates Inc., 2013, pp. 3111–3119.

[24] 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).* Melbourne, Australia, 2018, pp. 174–184. DOI: 10.18653/v1/P18-1017.

[25] S. M. Mohammad and P. D. Turney. “Crowdsourcing a Word–emotion Association Lexicon”. In: *Computational intelligence* 29.3 (2013), pp. 436–465. DOI: 10.1111/j.1467-8640.2012.00460.x.

[26] A. Moors, J. De Houwer, D. Hermans, S. Wanmaker, K. Van Schie, A.-L. Van Harmelen, M. De Schryver, J. De Winne, and M. Brysbaert. “Norms of Valence, Arousal, Dominance, and Age of Acquisition for 4,300 Dutch Words”. In: *Behavior research methods* 45 (2013), pp. 169–177. DOI: 10.3758/s13428-012-0243-8.

[27] E. T. Nalisnick and H. S. Baird. “Character-to-character Sentiment Analysis in Shakespeare's Plays”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* 2013, pp. 479–483.

[28] E. Öhman. “The validity of lexicon-based sentiment analysis in interdisciplinary research”. In: *Proceedings of the workshop on natural language processing for digital humanities.* NIT Silchar, India, 2021, pp. 7–12.

[29] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. 47. University of Illinois press, 1957.

[30] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds. “The Emotional Arcs of Stories are Dominated by Six Basic Shapes”. In: *EPJ data science* 5.1 (2016), pp. 1–12. DOI: 10.1140/epjds/s13688-016-0093-1.

[31] S. Rebora. “Sentiment Analysis in Literary Studies. A Critical Survey”. In: *DHQ: Digital Humanities Quarterly* 17.3 (2023).

[32] S. Rebora et al. “Shared Emotions in Reading Pirandello. An Experiment with Sentiment Analysis”. In: *Marras, C., Passarotti, M., Franzini, G., and Litta, E.(eds), Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive'per l'Informatica Umanistica. Università Cattolica del Sacro Cuore, Milano (2020)* (2020), pp. 216–221.

[33] R. Rehurek and P. Sojka. “Gensim–Python Framework for Vector Space Modelling”. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).

[34] S. Rosenthal, N. Farra, and P. Nakov. “SemEval-2017 task 4: Sentiment Analysis in Twitter”. In: *arXiv preprint arXiv:1912.00741* (2019). DOI: 10.18653/v1/S17-2088.

[35] S. W. Stirman and J. W. Pennebaker. “Word Use in the Poetry of Suicidal and Nonsuicidal Poets”. In: *Psychosomatic medicine* 63.4 (2001), pp. 517–522. DOI: 10.1097/00006842-2001 07000-00001.

[36] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT press, 1966.

[37] D. Teodorescu and S. Mohammad. “Evaluating Emotion Arcs across Languages: Bridging the Global Divide in Sentiment Analysis”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore, 2023, pp. 4124–4137. DOI: 10.18653/v1/2023 .findings-emnlp.271.

[38] W. Van Atteveldt, M. A. Van der Velden, and M. Boukes. “The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-coding, Dictionary approaches, and Machine Learning Algorithms”. In: *Communication Methods and Measures* 15.2 (2021), pp. 121–140. DOI: 10.1080/19312458.2020.1869198.

[39] L. Van Wissen and P. Boot. “An Electronic Translation of the LIWC Dictionary into Dutch”. In: *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing. Leiden, The Netherlands, 2017, pp. 703–715.

[40] A. B. Warriner, V. Kuperman, and M. Brysbaert. “Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas”. In: *Behavior research methods* 45 (2013), pp. 1191–1207. DOI: 10.3758/s13428-012-0314-x.

[41] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam. “A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.11 (2022), pp. 11019–11038. DOI: https://doi.ieeecomputersociety.org/10 .1109/TKDE.2022.3230975.

A. Appendix

A.1. Tools

A.2. Moors: Norms of valence, (...) for 4,300 Dutch words

The Moors approach is based on the Moors et al. article [26] discussed in the text. The valences reported by Moors vary from 1 (lowest) to 7 (highest).

A.2.1. Pattern: A Subjectivity Lexicon for Dutch Adjectives

We use the De Smedt and Daelemans subjectivity dictionary discussed in the text [8]. For the dictionary comparison, if a word occurs in the dictionary multiple times (because of multiple word senses) we take the average valence. For the computation of the fragment valence, we do not use the dictionary directly, but apply the Pattern toolset to the unlemmatised text fragment. Pattern does not just look at individual words, but uses some aspects of the context, such as the presence of intensifying adverbs ('awfully good').

A.2.2. LiLaH: The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene

The LiLaH dictionary [22] is a manually corrected version of an automatic translation of the NRC emotion lexicon [25] for three languages. It assigns words to positive or negative sentiment (+1 or -1) as well as to specific emotions. For Dutch, however, only the sentiment values are available. It contains 5746 Dutch words, 2519 are positive, 3431 are negative and 204 are both positive and negative. In the computation of the fragment valence, if a word is both positive and negative, we count its value as 0.

A.2.3. LIWC: Linguistic Inquiry and Word Count

LIWC is a general-purpose tool for text analysis created by psychologist James Pennebaker and colleagues. Its latest version is LIWC 2022 [5]. Underlying the tool is an English-language dictionary that assigns words to (multiple) categories, including categories for positive and negative emotion. There exist Dutch translations for the 2007 [4] and 2015 [39] versions of LIWC. The translation of the LIWC 2007 dictionary is a manual translation. It includes wildcards. The translation of LIWC 2015 is an automatic translation that resolved wildcards.

What LIWC reports is the relative frequency in a text of words in the various categories. We compute the relative frequencies using the LIWCTools python package [3]. For computation of the fragment valence we subtract the relative frequency of negative emotion from the relative frequency of positive emotion.

A.2.4. SentiArt: a word-embedding based computation

For SentiArt, we use a word2vec-computed [23] word embedding. We used the gensim package [33] to do the computation; we pre-tokenized, lemmatized, and lowercase the text with a window size of 8 words and only counted lemmas that appear in the corpus 5 or more times.

Table A.1
SentiArt valence computation options

Applied when computing valence of	Option
word	To compute or not compute the centroids (average value) for the positive words and the negative seed words before we compute and subtract the similarities.
word	To use a list of noun-only seed words or to include also corresponding adjectives (and one verb, where there is no corresponding adjective).
fragment	To use or not to use the lens method [18, p. 22], which excludes the second and third quartiles of the valence distribution from the computation.
fragment	To exclude or not punctuation characters.
fragment	To exclude or not function words. We use the function words as defined by Dutch LIWC 2007.

The texts that we used for the word-embedding are the full texts of 13,210 novels taken from a larger collection of 18,467 books in Dutch.

We use a number of different options in the SentiArt computation of the word and the fragment valences (see table A.1). Table A.2 gives the seed words that we use for computing the SentiArt valence. The fact that we included among the SentiArt options the possibilities to exclude punctuation and/or function words is the result of preliminary testing. We saw in these tests that punctuation and function word valences had a sizable effect on the SentiArt-assigned fragment valences, even under the 'lens' condition. E.g. the comma, the full stop and the indefinite article 'een' ('a') all have a SentiArt valence in the upper quartile of the distribution.

After computing the four SentiArt valence dictionaries, we look at their distributions. We also list the 50 words with the highest and lowest valence, in order to check whether the computation makes sense.

A.2.5. Transformer-based models

As mentioned, we use two transformer-based models on the fragments. One model is the cardiffnlp/twitter-xlm-roberta-base-sentiment model [1]. This is a multilingual model, trained on tweets. It does not assign a continuous valence value, but classifies a text as positive, negative or neutral.[2] The other model is robbert-v2-dutch-base-hebban-reviews5. It is a model trained on Dutch book reviews from the book discussion site Hebban, and aims to predict the rating associated with the review.[3] For both tools, the text types that they were trained on are very different from the book fragments that we will use them on. For this reason, we did not expect that they would agree with the other tools, but were willing to be surprised.

[2] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment.

[3] https://huggingface.co/BramVanroy/robbert-v2-dutch-base-hebban-reviews5.

Table A.2
Valence seed words

Condition	Positive	Negative
Basic	tevredenheid (contentment) blijdschap (joy) genot (pleasure) trots (pride) opluchting (relief) voldoening (satisfaction) verrassing (surprise)	walging (disgust) verlegenheid (shyness) angst (anxiety) verdriet (sadness) schaamte (shame)
Extended	*Basic seed words &* tevreden (contented) blij (glad) genieten (enjoy) trots (proud) opgelucht (relieved) voldaan (satisfied) verrast (surprised)	walgend (disgusted) verlegen (shy) angstig (anxious) verdrietig (sad) beschaamd (ashamed)

A.3. Computing the SentiArt valences

We computed four SentiArt valences, as described above. Here we used only the 17,306 lemmas that occur in the novel fragments that we will analyse.

As a first check of the results, we looked at the 50 words with highest or lowest valence for each of the computations. The words with the reportedly lowest valence are for all four computations indeed words that describe very unpleasant things. As an example, here are (English translations of) the 10 words with lowest valence for the centroid - nouns and adjectives condition: fear, distraught, anger, anxious, rage, shame, misunderstood, powerlessness, anger, confounded. For the words with highest valence, the picture is somewhat different. There are many words that no doubt represent a positive evaluation (excellent, fantastic, great), but there also appears a curious group of words that seem somehow related to hospitality, such as dinner, hostess, sommelier, catering, service and culinary; words that certainly are far removed from the positive seed words that went into the process.

Next we look at the distribution of the computed valences. Figure A.1 shows that the centroid-based computations, and especially the one with nouns and adjectives as seed words, have a somewhat wider distribution. That seems an attractive property, as it provides stronger distinctive power to the valence assignment.

The correlations between the SentiArt valences are very strong (Table A.3). We see there is a 2 to 3 percent disagreement between the centroid and non-centroid versions, and a 6 to 7 percent disagreement between the nouns versus the nouns and adjectives seed words.

For each dictionary pair, we selected 10 words that get a high valence rating in one dictionary but a low rating in another. In none of the pairs, this created word lists where intuitively we would consider one of the dictionaries wrong, except for the pair centroid and nouns / noncentroid and nouns. Here, the words that were rated high in the centroid but low in the

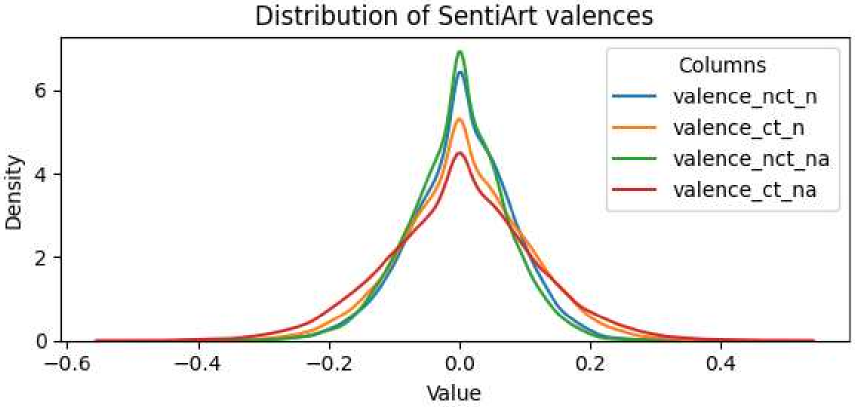


Figure A.1: Distribution of the four SentiArt valences. ‘ct’: centroid, ‘nct’: non-centroid, ‘n’: noun labels, ‘na’: noun and adjective labels.

Table A.3
Pearson correlations between the four SentiArt dictionaries. ‘ct’: centroid, ‘nct’: non-centroid, ‘n’: noun labels, ‘na’: noun and adjective labels.

	valence_nct_n	valence_ct_n	valence_nct_na	valence_ct_na
valence_nct_n	1.00			
valence_ct_n	0.97	1.00		
valence_nct_na	0.91	0.87	1.00	
valence_ct_na	0.94	0.93	0.98	1.00

non-centroid condition were all obviously positive: humour, eagerness, liveliness, surrender, delight, self-assurance, cheerfulness, approval, passion, lust. This provides another argument in favour of the centroid-based computation.

As explained in the previous section, for SentiArt we have five binary options, and therefore 32 different results. In initial testing, it appeared that the computation without centroid consistently led to results with lower correlations to the other tools than the computation with centroid. We dropped the computation without centroid from further consideration. From the remaining sixteen SentiArt valences, the best correlation with the other tools was reached with the options lens, just the original (noun) labels, and not considering punctuation and stop words (see Figure A.2 for Pearson correlation with the other dictionary-based tools.). We continue with this SentiArt valence, which in the rest of the paper we just call SentiArt.

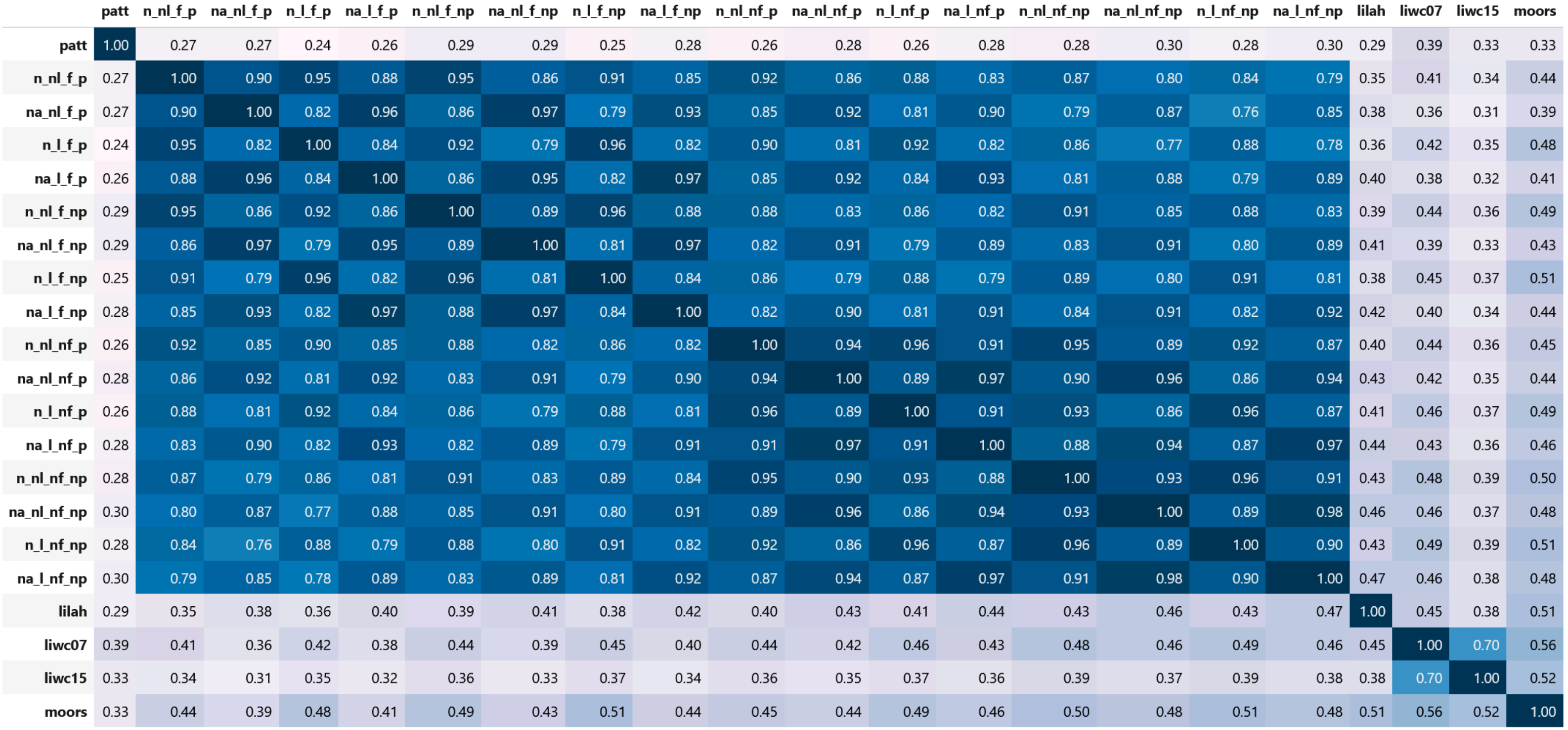
A.4. Other tables and figures

Table A.4
Overlap in words between the tools' dictionaries. A limitation to take into account: in LIWC07, some terms include wildcards. The number of words that it covers are therefore larger than the number reported here. The computation of the overlap does not take into account the wildcards.

tool 1	tool 2	in 1	in 1 not in 2	in 1 and 2	in 2 not in 1	in 2
sentiart	lilah	677636	672438	5198	548	5746
	liwc07	677636	676548	1088	1322	2410
	moors	677636	673357	4279	20	4299
	patt	677636	674413	3223	81	3304
lilah	liwc07	5746	5326	420	1990	2410
	moors	5746	4265	1481	2818	4299
	patt	5746	4842	904	2400	3304
liwc07	moors	2410	2076	334	3965	4299
	patt	2410	2202	208	3096	3304
moors	patt	4299	3573	726	2578	3304

Table A.5
Apparent errors in valence assignment (selection).

Tool	Unexpectedly positive	Unexpectedly negative
lilah	bombastic authoritarian	serious disinterested youthful warm
sentiart	fake sorrowful rigid boring disdainful sucky unattractive	mindful forgive kiss compassion loved ode innocence



	patt	n_nl_f_p	na_nl_f_p	n_l_f_p	na_l_f_p	n_nl_f_np	na_nl_f_np	n_l_f_np	na_l_f_np	n_nl_nf_p	na_nl_nf_p	n_l_nf_p	na_l_nf_p	n_nl_nf_np	na_nl_nf_np	n_l_nf_np	na_l_nf_np	lilah	liwc07	liwc15	moors
patt	1.00	0.27	0.27	0.24	0.26	0.29	0.29	0.25	0.28	0.26	0.28	0.26	0.28	0.28	0.30	0.28	0.30	0.29	0.39	0.33	0.33
n_nl_f_p	0.27	1.00	0.90	0.95	0.88	0.95	0.86	0.91	0.85	0.92	0.86	0.88	0.83	0.87	0.80	0.84	0.79	0.35	0.41	0.34	0.44
na_nl_f_p	0.27	0.90	1.00	0.82	0.96	0.86	0.97	0.79	0.93	0.85	0.92	0.81	0.90	0.79	0.87	0.76	0.85	0.38	0.36	0.31	0.39
n_l_f_p	0.24	0.95	0.82	1.00	0.84	0.92	0.79	0.96	0.82	0.90	0.81	0.92	0.82	0.86	0.77	0.88	0.78	0.36	0.42	0.35	0.48
na_l_f_p	0.26	0.88	0.96	0.84	1.00	0.86	0.95	0.82	0.97	0.85	0.92	0.84	0.93	0.81	0.88	0.79	0.89	0.40	0.38	0.32	0.41
n_nl_f_np	0.29	0.95	0.86	0.92	0.86	1.00	0.89	0.96	0.88	0.88	0.83	0.86	0.82	0.91	0.85	0.88	0.83	0.39	0.44	0.36	0.49
na_nl_f_np	0.29	0.86	0.97	0.79	0.95	0.89	1.00	0.81	0.97	0.82	0.91	0.79	0.89	0.83	0.91	0.80	0.89	0.41	0.39	0.33	0.43
n_l_f_np	0.25	0.91	0.79	0.96	0.82	0.96	0.81	1.00	0.84	0.86	0.79	0.88	0.79	0.89	0.80	0.91	0.81	0.38	0.45	0.37	0.51
na_l_f_np	0.28	0.85	0.93	0.82	0.97	0.88	0.97	0.84	1.00	0.82	0.90	0.81	0.91	0.84	0.91	0.82	0.92	0.42	0.40	0.34	0.44
n_nl_nf_p	0.26	0.92	0.85	0.90	0.85	0.88	0.82	0.86	0.82	1.00	0.94	0.96	0.91	0.95	0.89	0.92	0.87	0.40	0.44	0.36	0.45
na_nl_nf_p	0.28	0.86	0.92	0.81	0.92	0.83	0.91	0.79	0.90	0.94	1.00	0.89	0.97	0.90	0.96	0.86	0.94	0.43	0.42	0.35	0.44
n_l_nf_p	0.26	0.88	0.81	0.92	0.84	0.86	0.79	0.88	0.81	0.96	0.89	1.00	0.91	0.93	0.86	0.96	0.87	0.41	0.46	0.37	0.49
na_l_nf_p	0.28	0.83	0.90	0.82	0.93	0.82	0.89	0.79	0.91	0.91	0.97	0.91	1.00	0.88	0.94	0.87	0.97	0.44	0.43	0.36	0.46
n_nl_nf_np	0.28	0.87	0.79	0.86	0.81	0.91	0.83	0.89	0.84	0.95	0.90	0.93	0.88	1.00	0.93	0.96	0.91	0.43	0.48	0.39	0.50
na_nl_nf_np	0.30	0.80	0.87	0.77	0.88	0.85	0.91	0.80	0.91	0.89	0.96	0.86	0.94	0.93	1.00	0.89	0.98	0.46	0.46	0.37	0.48
n_l_nf_np	0.28	0.84	0.76	0.88	0.79	0.88	0.80	0.91	0.82	0.92	0.86	0.96	0.87	0.96	0.89	1.00	0.90	0.43	0.49	0.39	0.51
na_l_nf_np	0.30	0.79	0.85	0.78	0.89	0.83	0.89	0.81	0.92	0.87	0.94	0.87	0.97	0.91	0.98	0.90	1.00	0.47	0.46	0.38	0.48
lilah	0.29	0.35	0.38	0.36	0.40	0.39	0.41	0.38	0.42	0.40	0.43	0.41	0.44	0.43	0.46	0.43	0.47	1.00	0.45	0.38	0.51
liwc07	0.39	0.41	0.36	0.42	0.38	0.44	0.39	0.45	0.40	0.44	0.42	0.46	0.43	0.48	0.46	0.49	0.46	0.45	1.00	0.70	0.56
liwc15	0.33	0.34	0.31	0.35	0.32	0.36	0.33	0.37	0.34	0.36	0.35	0.37	0.36	0.39	0.37	0.39	0.38	0.38	0.70	1.00	0.52
moors	0.33	0.44	0.39	0.48	0.41	0.49	0.43	0.51	0.44	0.45	0.44	0.49	0.46	0.50	0.48	0.51	0.48	0.51	0.56	0.52	1.00

Figure A.2: Pearson correlations of all valences excluding the ones from the Transformer-based models. The Sentiart valences are named based on the options used in their computation. ‘n’: noun labels, ‘na’: noun and adjective labels, ‘l’: lens method, ‘nl’: no lens, ‘f’: include function words, ‘nf’: no function words, ‘p’: include punctuation, ‘np’: no punctuation.