AI

MDPI

Article

Computing the Affective-Aesthetic Potential of Literary Texts

Arthur M. Jacobs [1,2,*] and Annette Kinder [3]

[1] Department of Experimental and Neurocognitive Psychology, Freie Universität Berlin/FUB, 14195 Berlin, Germany
[2] Center for Cognitive Neuroscience Berlin (CCNB), FUB, 14195 Berlin, Germany
[3] Department of Education and Psychology, FUB, 14195 Berlin, Germany; annette.kinder@fu-berlin.de
* Correspondence: ajacobs@zedat.fu-berlin.de

Received: 24 November 2019; Accepted: 27 December 2019; Published: 30 December 2019

Abstract: In this paper, we compute the affective-aesthetic potential (AAP) of literary texts by using a simple sentiment analysis tool called *SentiArt*. In contrast to other established tools, *SentiArt* is based on publicly available vector space models (VSMs) and requires no emotional dictionary, thus making it applicable in any language for which VSMs have been made available (>150 so far) and avoiding issues of low coverage. In a first study, the AAP values of all words of a widely used lexical databank for German were computed and the VSM's ability in representing concrete and more abstract semantic concepts was demonstrated. In a second study, *SentiArt* was used to predict ~2800 human word valence ratings and shown to have a high predictive accuracy ($R^2 > 0.5$, $p < 0.0001$). A third study tested the validity of *SentiArt* in predicting emotional states over (narrative) time using human liking ratings from reading a story. Again, the predictive accuracy was highly significant: $R^2_{adj} = 0.46$, $p < 0.0001$, establishing the *SentiArt* tool as a promising candidate for lexical sentiment analyses at both the micro- and macrolevels, i.e., short and long literary materials. Possibilities and limitations of lexical VSM-based sentiment analyses of diverse complex literary texts are discussed in the light of these results.

Keywords: sentiment analysis; SentiArt; computational poetics; affective-aesthetic potential; machine learning; digital humanities; neuroaesthetics; literary reading

1. Introduction

Emotion recognition is a vital aspect of daily human life, important for survival, social, or professional reasons. However, only very recently—in evolutionary terms—has it become a challenge to both human readers and computer algorithms to read out emotional information from (literary) texts, e.g., when using machine-learning-assisted sentiment analysis/SA tools. Perhaps more than other objects of culture, written texts can induce emotions, since narratives are inseparable from the emotional content of the plots [1,2]. These emotions or sentiments can determine the most ubiquitous and basic affective decision of daily life, namely deciding whether we like or dislike something/somebody [3,4]. What we read about something or somebody also can determine our behavior, e.g., choosing a movie, buying a book, or voting for someone.

Sentiment analysis (SA) can be defined as: 'the process of computationally identifying and categorizing opinions (According to Liu (2015) an opinion is a quintuple, e_i, a_{ij}, s_{ijkl}, h_k, t_l, where e_i is a named entity (e.g., Abraham), a_{ij} an aspect of e_i (e.g., a word or phrase expressing an aspect such as 'Abraham's son is sad'), s_{ijkl} is the sentiment on aspect a_i, (e.g., a valence value or a discrete emotion label such as 'sad'), h_k is the opinion holder, and t_l is the time of the opinion expressing event) expressed in a piece of text, especially in order to determine whether the writer's attitude

towards a particular topic, product, etc., is positive, negative, or neutral' (Oxford English dictionary: https://www.lexico.com/en/definition/sentiment_analysis). Although the majority of SAs today are applied to book or movie reviews with several gold standards allowing one to evaluate the SA tool's performance [5], an increasing number of studies applies SA to literature and poetry [6–22].

The standard and most straightforward SA approach nowadays in natural language processing/NLP, digital humanities, computational linguistics and stylistics, psychology, or neurocognitive poetics is the *lexical* one: It simplifies complex emotional information analysis to a vocabulary-based computation of the polarity, valence, or some other sentiment variable of single keywords contained in the sentences of the text. Following Miller's psycholinguistic doctrine [23], whoever wants to understand how larger text segments can induce emotional processes must start with those basic units at which all relevant processes and representations in language use come together: Single words [24]. Words, as has long been known [25,26] are embodied stimuli with the potential to elicit overt and covert sensorimotor and affective responses [27]. They even can 'stink', suggesting that the affective processes we experience when reading rely on the reuse of phylogenetically ancient brain structures that process basic emotions in other domains and species [28].

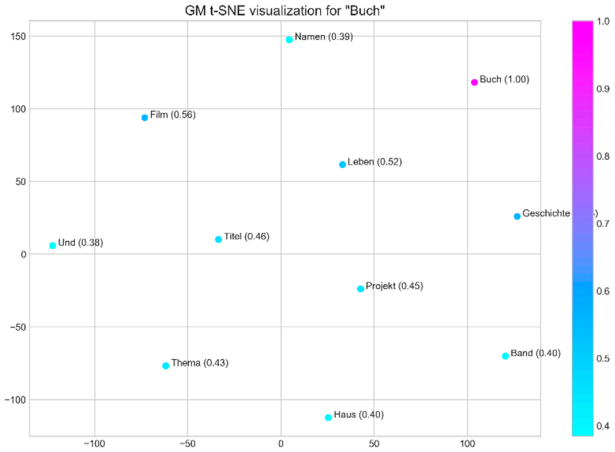
Regarding lexical SA, Bestgen's pioneering study [8] already suggested that lexical valence can predict the affective tones of sentences and entire texts quite well, and there is also a lot of recent evidence for the usefulness and empirical validity of the lexical approach to SA using different tools like *VADER* [5], *HU-LIU* [29], or *SentiArt* [13]. Like most other tools in the field, the former two are both based on word lists containing human rating data, i.e., what is sometimes called emotional dictionaries or prior-polarity lexicons [30]: *Vader* uses ~7.500 entries https://www.kaggle.com/nltkdata/vader-lexicon, and *Hu-Liu* ~6.800; (https://www.cs.uic.edu/~{}liub/FBS/sentiment-analysis.html#lexicon). In contrast, *SentiArt* uses an unsupervised learning approach introduced by Turney [31], which is based on vector space models (VSMs) and a label list representing prototypes of positive and negative *semantic orientation* or *emotional valence*, such as the labels GOOD, NICE vs. BAD, NASTY [32]. Optimally, word list-based methods should cover each (content) word—or at least a maximum—in the test texts to be 'sentiment analyzed' in order to augment both the reliability and validity of the tool. Practically, however, such tools often run into problems for several reasons. First, when dealing with highly literary or ancient text materials, the word lists overall coverage or hit rate can sink below 50% making the sentiment analysis unreliable. An example is given in Study 3 below, in which two word-list-based methods that are compared with *SentiArt* yield suboptimal results due to their low coverage when applied to a classical text in German, E.T.A. Hoffmann's (1816) *The Sandman.* Secondly, if there are no or only limited word lists available in the language of a researcher's country, simply translating existing English lists into that language without empirical cross-validation is problematic [33] since sensitivity to emotional content varies across languages, which differ considerably in their emotion vocabularies [34–36]. Collecting human rating data to create new word lists in other languages or to enlarge existing English ones is costly, but most importantly, there are serious methodological and epistemological issues about the reliability and validity of human sentiment ratings when they are turned from a dependent variable (i.e., a 'subjective' behavioral measure in response to a stimulus) into an independent variable (i.e., an 'objective' predictor of say the positivity of a text [37,38]).

The big advantage of VSM-based methods like *SentiArt* is that they avoid these problems: (i) They require no word lists based on human ratings; (ii) thanks to the public availability of VSMs in >150 languages (https://fasttext.cc/docs/en/pretrained-vectors.html) they can be applied to a multitude of texts from different countries even in special dialects; and (iii) by creating task- or domain-specific VSMs, they can be flexibly adapted to different research purposes, e.g., predicting human behavior of participants reading children books or Shakespeare sonnets [39,40]. The next section describes the workflow and exact procedure of *SentiArt*.

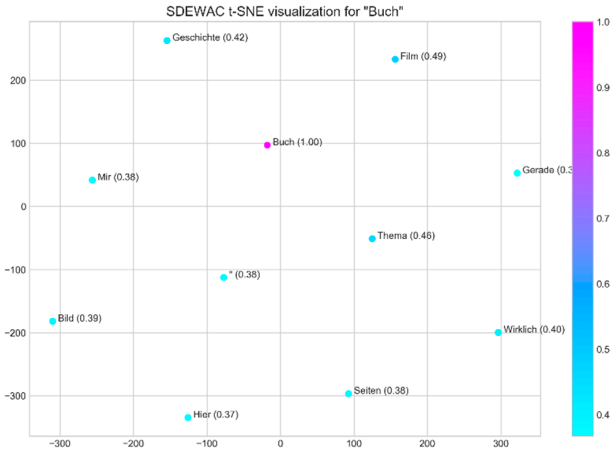
2. SentiArt

Computing the valence, arousal, or affective-aesthetic potential (AAP) of a word, sentence, or larger text segment with SentiArt requires only two things: (i) The standard *SentiArt* table for English (available from the first author via email) that provides valence, arousal, and AAP values for ~1 million English words from the publicly available *wiki.en* VSM, and (ii) a table providing the words—and optimally their corresponding part of speech tags—of the test text (provided by users). Using a simple table update procedure (e.g., via Excel or a python routine), users can read out say the valence value for each selected word in the test text from the *SentiArt* table and then compute the sums and averages per sentence or chapter.

A novelty of *SentiArt* thus lies in the fact that it can very easily be applied by non-expert users to use a general 2d emotion potential space (valence X arousal) as a reference for quantifying the emotion potential of a test text by localizing the position of its words in this standardized space (The words of a test text can be either all types (i.e., unique words) including function words or a reduced set of content words, of lemmas, or content word lemmas, depending on the researcher's interest and choice). A link to Figure 1 from Jacobs [13] illustrates the 2d space and the basic procedure applied for creating it (https://www.frontiersin.org/files/Articles/441916/frobt-06-00053-HTML/image_m/frobt-06-00053-g001.jpg). Typical users only need the *SentiArt* table, though, and do not follow the workflow described in Figure 1a in the link above. For many applications using standard English texts, this procedure will be sufficient to run a sentiment analysis.

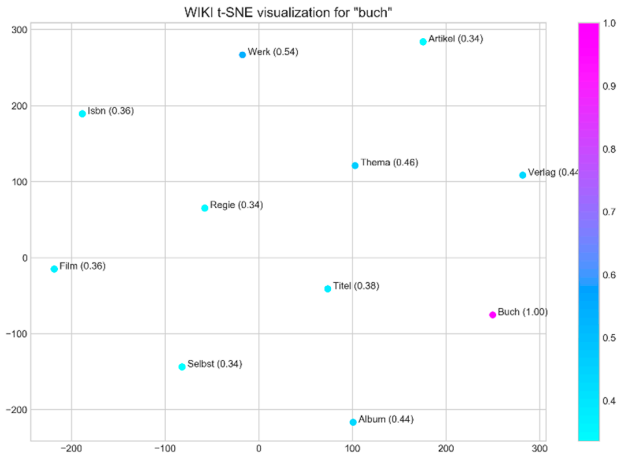


(**a**)



(**b**)

Figure 1. *Cont.*



(**c**)

Figure 1. (**a**–**c**) t-distributed stochastic neighbor embedding (tsne) semantic neighborhood representation for word BOOK ('Buch') in three VSMs with the 10 closest semantic neighbors.

However, as mentioned above, some users might want to create their own task-specific VSM, e.g., because they have reasons to think that the wiki.en VSM is not the optimal one for their test texts at hand. In this case, the workflow for applying *SentiArt* is as follows:

1. Selection and evaluation of an appropriate VSM, e.g., one can use the procedure described on the *fasttext* homepage (https://fasttext.cc/docs/en/pretrained-vectors.html) to directly download the (German) VSM called 'wiki.de.vec' providing 300d sublexical vectors for each of >2 million words (e.g., in the original uncleaned version [41]).
2. Selection and evaluation of an appropriate label list, e.g., for valence, one could use the model-based label lists empirically validated [13,42].
3. Computation of AAP and evaluation of predictive accuracy, i.e. cross-validation with empirical data (e.g., human ratings)

Each of these three steps involves multiple choices, which influence the results of the SA and will be described in detail below.

2.1. The Present Study

In this paper, we test the accuracy and validity of *SentiArt* and its underlying computational tools by applying it to the different test materials and discuss possibilities and limitations of lexical SA of complex literary texts.

Study 1. Selecting and evaluating the VSMs

Two crucial ingredients of VSM-based SA are the training corpus and VSM. The VSM ($M = R^{v \times d}$, where v is the size of the vocabulary and d the dimensionality) is always based on a training corpus and the choice of the latter will affect the quality and utility of the VSM for the SA purposes at hand. For example, the publicly available VSM called 'german.model/GM with $GM = R^{610k \times 300}$; (https://devmount.github.io/GermanWordEmbeddings/) was trained with the 'word2vec' algorithm [43] on the German Wikipedia and news articles of a single day in a specific year (15 May 2015). Thus, the VSM incorporates choices regarding the size, representativeness, or specificity of the training corpus, all of which will influence the quality and validity of the VSM used to compute the semantic relatedness values, which are crucial for establishing the 2D emotion potential space and thus the results of the SA (for an overview of relevant training corpora, see [12]).

For the present purpose, we chose the widely used, publicly available, and ecologically valid *subtlex* database for German [44] as a reference lexicon, which allowed us an evaluative comparison between three VSMs using an identical set of items. The ~120k words of *subtlex* overlap sufficiently with those of a large range of both non-literary and literary texts to obtain stable SA results. We further chose three publicly available German VSMs as a basis for 'sentiarting' the ~120k words of subtlex (i.e., assigning VSM-based valence and AAP values to each word), which was then used to predict the valence and AAP values of the words of our test texts. Each VSM was evaluated using a face-validity approach based on the t-distributed stochastic neighbor embedding (tsne) algorithm [45], as well as a cross-validation procedure using human valence rating data from the *Berlin Affective Word List* (BAWL) [46,47]. Table 1 summarizes the data for the three VSMs.

Table 1. VSM specifications for the present study.

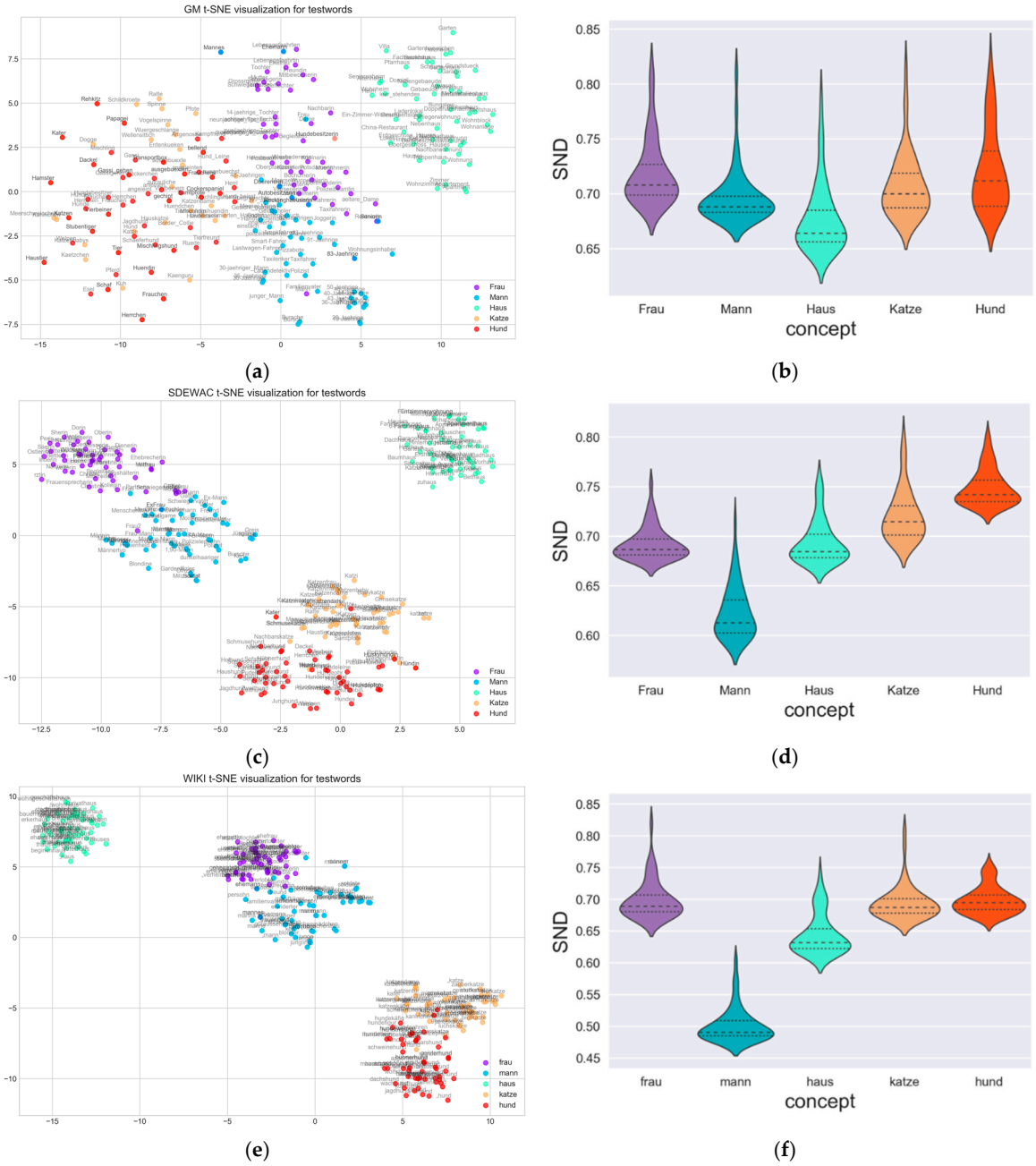
VSM	Size (Cleaned [1]), DIMENSIONALITY	Overlap with Subtlex (in Number of Unique Words)	Original Training Corpus
German.model/GM https://devmount.github.io/GermanWordEmbeddings/	608.130 (384.183) 300d	86.049	German Wikipedia and news articles (15th May 2015)
Sentence Dewac/SDEWAC https://www.ims.uni-stuttgart.de/en/	1.592.753 (1.354.303) 300d	116.497	unspecified German texts from the web [48]
Wiki.de/WIKI https://fasttext.cc/docs/en/pretrained-vectors.html	2.275.233 (2.133.318) 300d	114.198	unspecified German texts from wikipedia

[1] The cleaning procedure simply deleted all words containing non-alphabetic characters.

2.2. VSM Evaluation

When human rating or other empirical data are available, the validity of a VSM and the VSM-based SA can straightforwardly be cross-validated [13,49–52]. If such data are not available, face validity tests, e.g., using semantic arithmetic experiments, are a viable option. The model evaluation tests proposed for the 'german.model' are exemplary in this regard, including multiple systematic semantic arithmetic and syntactic tests.

Figures 1–3 compare the face validity of the three VSMs sketched in Table 1 using the tsne algorithm. The idea behind this tsne-based evaluation is that if the VSM is any good for the present purposes, concepts obviously related to each other, such as the emotionally rather neutral CATS and DOGS ('Katze', 'Hund'; Figure 2), or the more emotionally valenced concepts in Figure 3 (e.g., DISGUST/'Ekel') should be separated but relatively close in semantic space, while a concept like 'house' should be clearly apart. This is indeed the case for all three VSMs in German.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 2. (**a**–**f**) tsne semantic neighborhood representation and violin plots of semantic neighborhood density (SND, i.e., the average cosine of the target item with its 50 nearest neighbors; the violin plots show the distributions together with quantiles) for five test words (WOMAN, MAN, HOUSE, CAT, DOG) in three vector space models (VSMs).

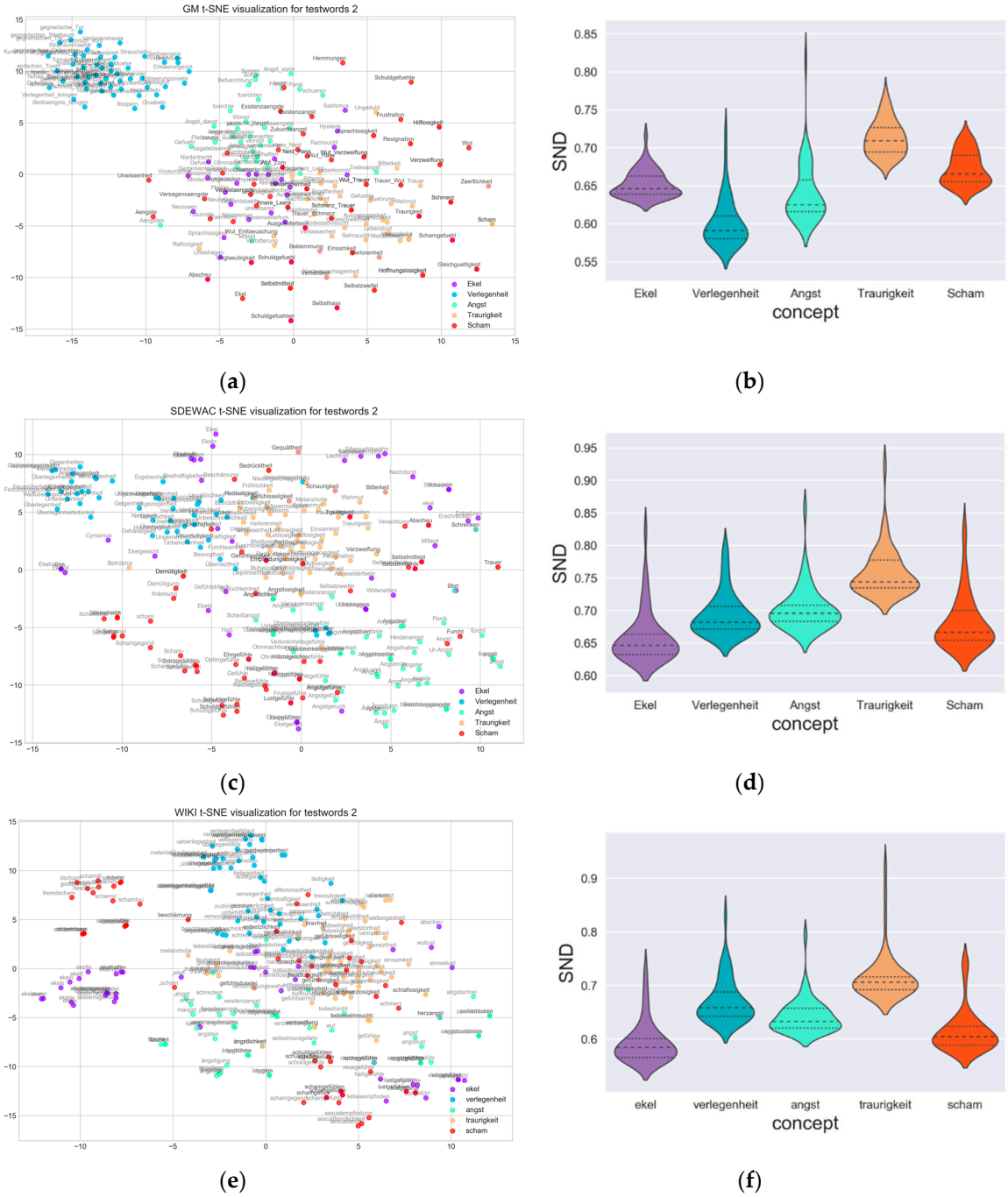


Figure 3. (**a**–**f**) tsne semantic neighborhood representation and violin plots of semantic neighborhood density (SND) for five test words (DISGUST, EMBARRASSMENT, FEAR, SADNESS, SHAME) in three VSMs with their 50 closest semantic neighbors.

The results in these three Figures can be summarized as follows. First, the data show that, as expected, overall each VSM generates distinct semantic neighborhoods as represented by the tsne method. In Figure 1a, GM produces STORY ('Geschichte') and MOVIE ('Film') as the closest semantic neighbors of the target word BOOK ('Buch'), for SDEWAC it is MOVIE ('Film') and THEME ('Thema'), while WIKI produces OPUS ('Werk') and THEME ('Thema'). The data for the list of rather neutral words in Figure 2 (WOMAN, MAN, HOUSE, CAT, DOG) suggest that with increasing VSM size, the concepts corresponding to these words become better separated in the 2D computational semantic space. Thus, while in GM (Figure 2a,b) the concepts CAT ('Katze', yellow dots) and DOG

('Hund', red dots) still widely overlap and are close to WOMAN ('Frau', magenta dots) and MAN ('Mann', cyan dots), in both SDEWAC (Figure 2c,d) and WIKI (Figure 2e,f), they are clearly apart from each other and from the concept HOUSE ('Haus', green dots). The violin plots show semantic neighborhood density (snd) for each concept as quantified by the average cosine of the target item with its N nearest neighbors (where N was set to 50 here) [53].

Finally, the emotional word list in Figure 3, which corresponds to the five negative labels from the 'Ekman99' model [42], i.e., DISGUST ('Ekel', magenta), EMBARRASMENT ('Verlegenheit', cyan), FEAR ('Angst', green), SADNESS ('Traurigkeit', orange), and SHAME ('Scham', red). The data show that the conceptual overlap is much larger than for the neutral words in Figure 2, a finding that can be expected given the relatively abstract nature of emotion terms compared to concrete categories like DOG. As can also be seen in Figure 3b,d,f, the three VSM produce different snd values with different distributional shapes, although the concept SADNESS (orange) appears to be the 'clearest' (i.e., highest snd) in all three VSMs. On the basis of these descriptive data in Figures 1–3, one can expect notable differences between the three VSMs with regard to predictions concerning the valence and AAP of texts. The next study describes the computation of the AAP together with the evaluation of the label lists.

Study 2. Computation and validation of lexical valence and AAP values.

Each word of the *subtlex* database was 'sentiarted' as follows. Using the vectors from the three models summarized in Table 1, we computed the valence values based on the semantic relatedness (estimated via the cosine similarity between two vectors) between each word in subtlex and the theoretically motivated and empirically validated 'Ekman99 emotion labels' [42]. The model-based valence of a test word, $v(w)$, is computed according to Equation (1), i.e., as the difference between two average similarity values: First, the average similarity (s) between the test word (w), and the seven (m) positive emotion labels ($l_{pos\_1–7,}$ = CONTENTMENT, HAPPINESS, PLEASURE, PRIDE, RELIEF, SATISFACTION, SURPRISE), and, second, the average similarity between the test word and the five (n) negative emotion labels ($l_{neg\_1–5,}$ = DISGUST, EMBARRASMENT, FEAR, SADNESS, SHAME).

$$v(w) = \sum_{i=1}^{m} \mathrm{s}\left(\mathrm{w}, l_{pos_i}\right)/m - \sum_{i=1}^{n} \mathrm{s}\left(\mathrm{w}, l_{neg_i}\right)/n \quad (1)$$

The similarity between a word and a label, $\mathrm{s}(\mathrm{w},\mathrm{l})$, is computed by the cosine between the 300d vectors for word and label, as given by the VSM, shown in Equation (2). $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$.

$$\mathrm{s}(\mathrm{a},\mathrm{b}) = \frac{\sum_{i=1}^{300} A_i\, B_i}{\sqrt{\sum_{i=1}^{300} {A_i}^2}\sqrt{\sum_{i=1}^{300} {B_i}^2}} \quad (2)$$

where A_i and B_i are the vectors for word and label, respectively.

As an example, using the SDEWAC VSM, the computational valence, $v(w)$, for the theoretically most positive test word 'reizvoll' (APPEALING) in the *subtlex* database yielded an average similarity with the seven positive labels of 0.23 and an average similarity with the five negative labels of 0.19, resulting in a theoretical valence of 0.04.

The same procedure is applied for the computation of arousal and AAP values for each test word. For the latter ones, we used the extended 120 label list [12,48], which is given in the Appendices of both papers. All these values are summarized in an '.xlsx' table available via email (ajacobs@zedat.fu-berlin.de). The SDEWAC sheet, for example, gives the valence, arousal, and AAP values for each of ~120k words from the *subtlex* database that overlap with the SDEWAC words (the original German subtlex database has ~200k words, but here we only used those for which a spelling check had been made (~125k). ~115k of these words overlap with those in the wiki.de VSM, ~90k with the german.model, and ~120k with sdewac).

2.3. Label List Evaluation Predicting Human Valence Rating Data

Among the 12 models based on psychological emotion theories that Westbury [42] tested as candidate label lists, the 'Ekman99' [54] model with the above outlined 12 labels was the winner accounting for about 34% variance in the validation set of >10.000 human valence ratings database [42,55]. Here, we used publicly available German valence rating data from the BAWL [46,47], a very successful tool, which has been applied in >100 studies in different fields of research [10], to test the validity of the translated 12 'Ekman99' labels when used with the three VSM. It is worth noting that the same validation procedure should, in principle, be applied when using alternative word list based SA tools, i.e., before using them, one should test how well they predict human ratings from an independent data base, such as the BAWL or the one by Warriner [55]. As far as we can tell, such a cross-validation procedure is not yet standard practice, though.

The data in Figure 4 establish an interesting novel finding: The best predictor of human valence rating data, at least for the German BAWL, is not computational valence based on the 'Ekman99' label list, but AAP based on the AAP list [48], the latter accounting for more than twice the variance than the former. In a way, this is not astonishing, since the former uses only 12 labels and the latter uses 120 (60 positive and 60 negative items including almost all 'Ekman99' labels) thus making it much less context sensitive and more accurate. This is true for all three VSMs. The VSM yielding the best performance ($R^2 > 0.5$) is SDEWAC (middle panel), followed by WIKI and GM. Model performance can be increased to > 0.6 when reducing the rating data to those items which have the highest inter-rater agreement, e.g., for items with a standard deviation of ≤ 0.9. However, compared to the results of Westbury [42], an $R^2 > 0.5$ appears pretty good. Human valence ratings very likely are based, in a yet unknown part, on information retrieved from semantic memory which is of experiential/embodied origin, as opposed to distributional semantics [2,56]. The perhaps simplest assumption would be to consider this 'embodied part' to account for ~50% of the variance. If this would hold, accounting for ~50% of the variance by means of distributional semantic models seems very promising to us. Given these cross-validation results, we used this best-fitting VSM for all following SAs.

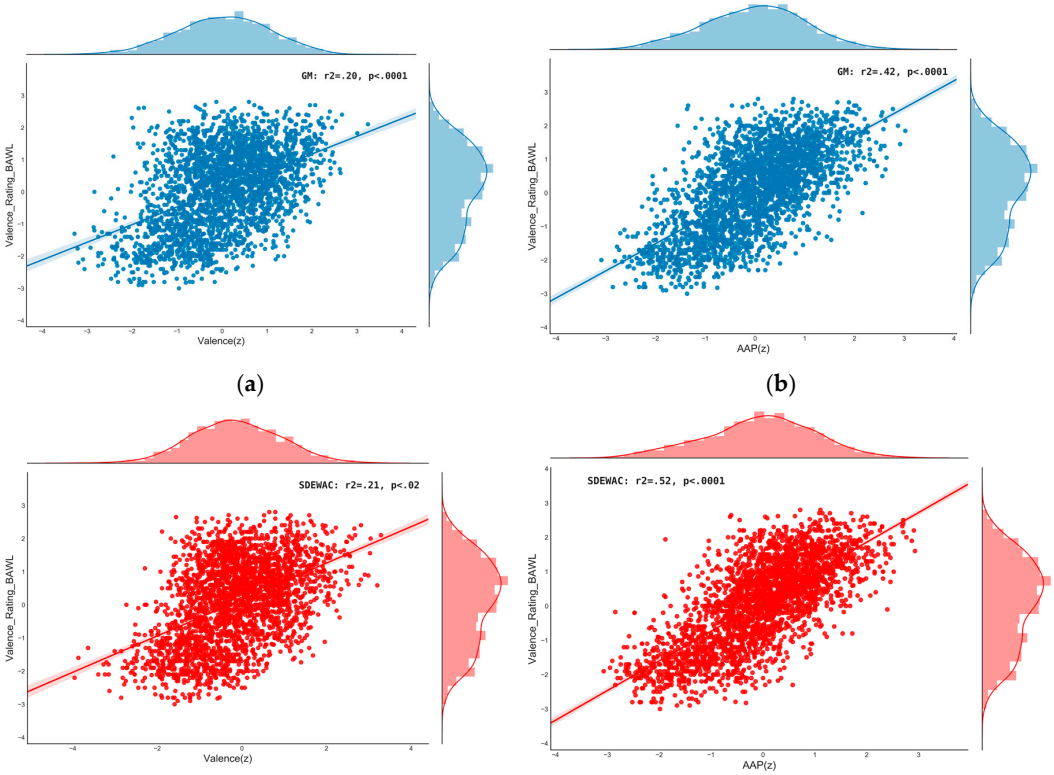


Figure 4. *Cont.*

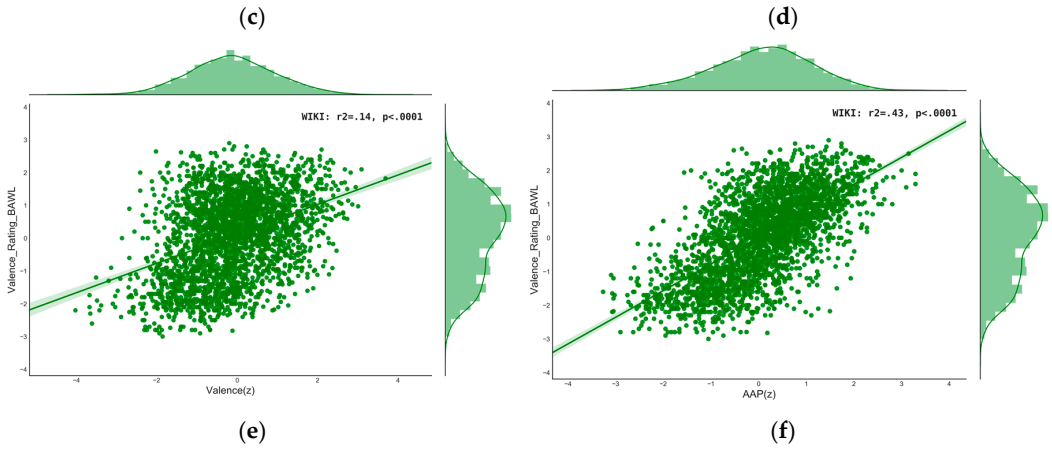


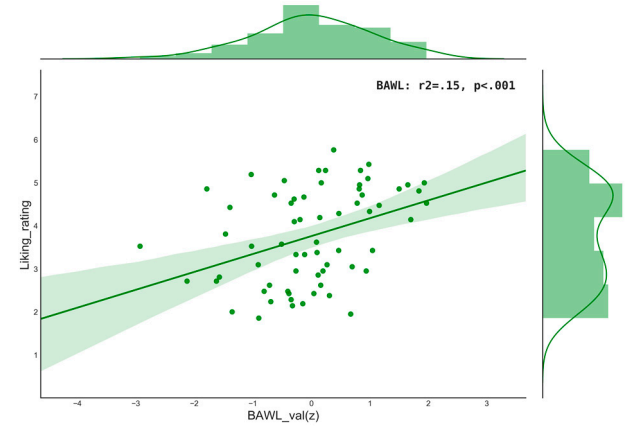
Figure 4. (**a**–**f**) Human valence ratings as predicted by the valence and affective-aesthetic potential (AAP) values computed with three different VSMs. Upper panel: German model (GM), Middle panel: SDEWAC, lower panel: WIKI.

Study 3. Predicting human liking ratings and emotional states over time with different AAP indices

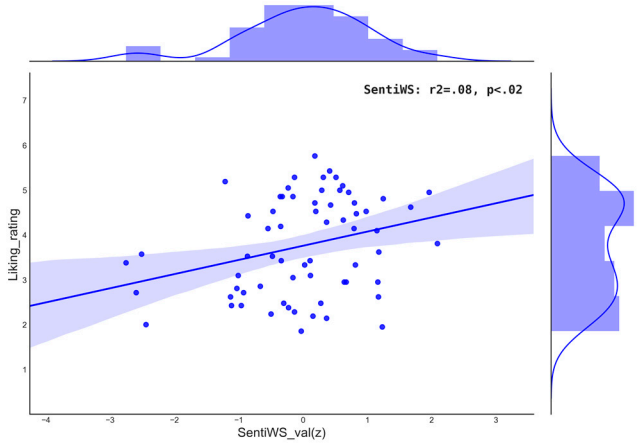
Using the look-up table approach the computation of the mean AAP of stories or book chapters is straightforward, coming down to simply cross-referencing a list of words representing a book chapter (or a sentence or paragraph) with the corresponding words in the 'sentiarted' subtlex table, as described above.

2.4. Evaluation of the AAP Construct Predicting Human Liking Ratings

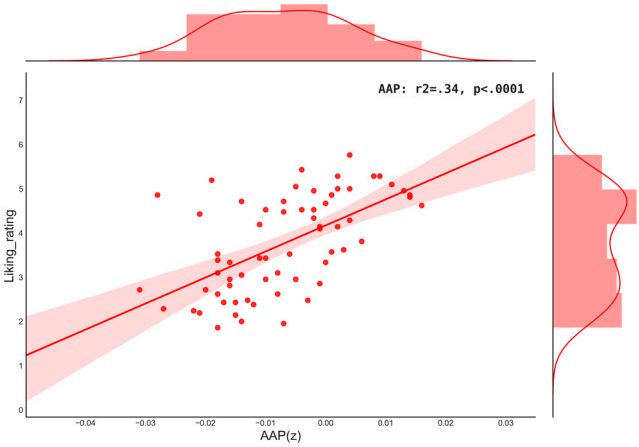
The superior predictive validity of the present AAP construct (over the valence construct) was established using human valence rating data (Figure 4). Here, we tested it a second time against human liking ratings from a reading study using E.T.A. Hoffmann's *The Sandman*—a prototypically uncanny narrative from 1816 representative of the 'black romantic' that evokes feelings of suspense and immersion in readers [14,57]. (In this study, 20 participants first read the story ,The Sandman' (divided into 65 segments of approximately equal length; M = 105.5 words; SD = 26.1 words) on paper in one go. Afterwards, they had to answer five comprehension questions to ensure that they had actually read and understood the story. Finally, the novella was returned to them and they rated each of the 65 sections separately on a computer on different scales (liking scale = 1–7). All ratings referred to subjects' experience during the first reading, which was explicitly pointed out to them. The entire experiment lasted between 90 and 140 min depending on the reader. The data were averaged across readers) For comparison, the liking ratings were predicted by three different predictors: (i) Empirically measured BAWL valence ratings (Figure 5a: Green), (ii) *SentiWS* values obtained from the German SA tool of Remus [58] (Figure 5b: Blue), and (iii) the present computational AAP values (Figure 5c: Red).



(**a**)



(**b**)



(**c**)

Figure 5. (**a**–**c**) Human liking ratings as predicted by three different dentiment analysis (SA) methods: (**a**) Empirically measured Berlin affective word list (BAWL) valence ratings, (**b**) *SentiWS* lexicon values, and (**c**) computational AAP values.

While all R^2 values are moderate, the model fits are highly significant and, most importantly, the fit for the two empirically derived predictors (BAWL and sentiWS ratings) is not better than the one for AAP, on the contrary. Presumably, part of the superior performance of the AAP method lies in the fact that the hit rate (content words only) of the other two is low, which makes their estimates unreliable (SentiWS ~ 15%, BAWL ~ 30%, AAP ~ 90%). Together with successful previous applications

of *SentiArt* in the prediction of word beauty ratings [48], or the classification of text segments from the Harry Potter books [13,59], these data establish the AAP, as computed by *SentiArt* with the SDEWAC VSM, as a viable alternative to using human rating data as predictors of *other* human rating data, an often costly and, in general, epistemologically questionable method [37,38].

2.5. Predicting Emotional States Over (Narrative) Time

One challenge addressed by the present research topic of this journal is defined by the fact that 'modeling and predicting the emotional state over time is not a trivial problem, because continuous data labeling is costly and not always feasible. This is a crucial issue in real-world applications, where the labeling of the features is sparse and eventually describes only the most prominent emotional events'. Stories or books are natural candidates for analyzing emotional states over time, since they offer the possibility to plot the emotion potential across different chapters or other units of narrative time. It should be noted though that in many cases, narrative time is not linear and thus cannot always be directly compared to the results provided by the present approach (We are grateful to an anonymous reviewer for mentioning this). However, for reasons of comparability, here we followed the standard procedure proposed by successful *macroanalytic* approaches for analyzing emotional time series of entire books, such as the 'hedonometer' [9] or the 'Syuzhet' package [15]. These methods aggregate lexical SA information across large units of texts (e.g., 10.000 words for the hedonometer in a linear way [60]).

While not all literary texts are equally well suited to such macroanalyses, the emotion potential method proposed in an earlier paper [12] can also be reliably applied to small text units, i.e., for *microanalyses*, such as a single Shakespeare sonnet (~115 words) or short text segments of ~100 words like the present ones from *The Sandman* (overall length ~7000 words). Such smaller texts are well suited to empirically cross-validate the theoretical predictions derived from SA tools by collecting (quasi-)continuous rating data. At least two such datasets have been examined in previous studies [57,61], which were not interested in computational SA though.

Using the above data from *The Sandman* study, next we tested the prediction of human liking ratings over time, i.e., the evolution of the story across the 65 segments, by different AAP indices. When readers judge the emotional content of subsequent coherent pieces of text, which kind of text features they really use for that complex decision still is a big open question both for NLP and neurocognitive poetics approaches [3,12]. Thus, focusing on the lexical level, we don't know yet whether readers take into account all words of the text or just a few key words, whether they pay attention to word forms (conjugation, inflection, derivation) and/or recurrent words (integrating word frequency information) or not. If they do not take into account each and every word, other open questions are whether content words count most (or exclusively), or whether words with extreme valence values weigh more. To shed some light on this issue, we tested several models here:

- Model A: All words, i.e., the AAP value for a text segment corresponded to the mean of the AAP values for all unique words (types) in the segment (which also occur in the Subtlex), including function words (N = 4723; mean $R^2_{adj} = 0.21$).
- Model L: All lemmata (N = 4685; mean $R^2_{adj} = 0.18$).
- Model C: All content words (N = 3080; mean $R^2_{adj} = 0.21$).
- Model CL: All content lemmata (N = 3044; mean $R^2_{adj} = 0.16$).

For each of the four models, we further computed several AAP indices: (i) Mean AAP (mean $R^2_{adj} = 0.32$), (ii) frequency-weighted mean AAP (mean $R^2_{adj} = 0.28$); (iii) lens mean AAP (mean $R^2_{adj} = 0.05$; The 'lens' option was proposed by Dodds [9] to obtain a strong signal by only keeping words residing in the tails of the valence distribution). Here we took all words into account for which AAP was <25% or >75% of the distribution), (iv) frequency-weighted lens mean AAP (mean $R^2_{adj} = 0.11$). The AAP values resulting from all $4 \times 4 = 16$ computing methods were then used as predictors of the human liking ratings in 16 linear regression models. The mean R^2 values for the

models and indices indicated above suggest that using lemmata or lens extremes did not help the present SA. The winning model C was based on the simple mean of the AAP values for all unique content words without frequency weighting ($R^2_{adj} = 0.34$), closely followed by the frequency-weighted variant ($R^2_{adj} = 0.33$). Interestingly, this suggests that, at least for the present short segments of a mystery story, readers seem to have focused on content words but not to have relied much on a cumulative AAP value, largely ignoring how often a given content word occurred. Of course, readers very likely also use inter- or supralexical information in their liking ratings of literary texts [3] thus explaining the moderate R^2 values, which leave about 70% of variance unaccounted for. Still, the data in Figure 6 look very promising in showing the potential of a purely lexical micro-SA for predicting emotional states over narrative time.

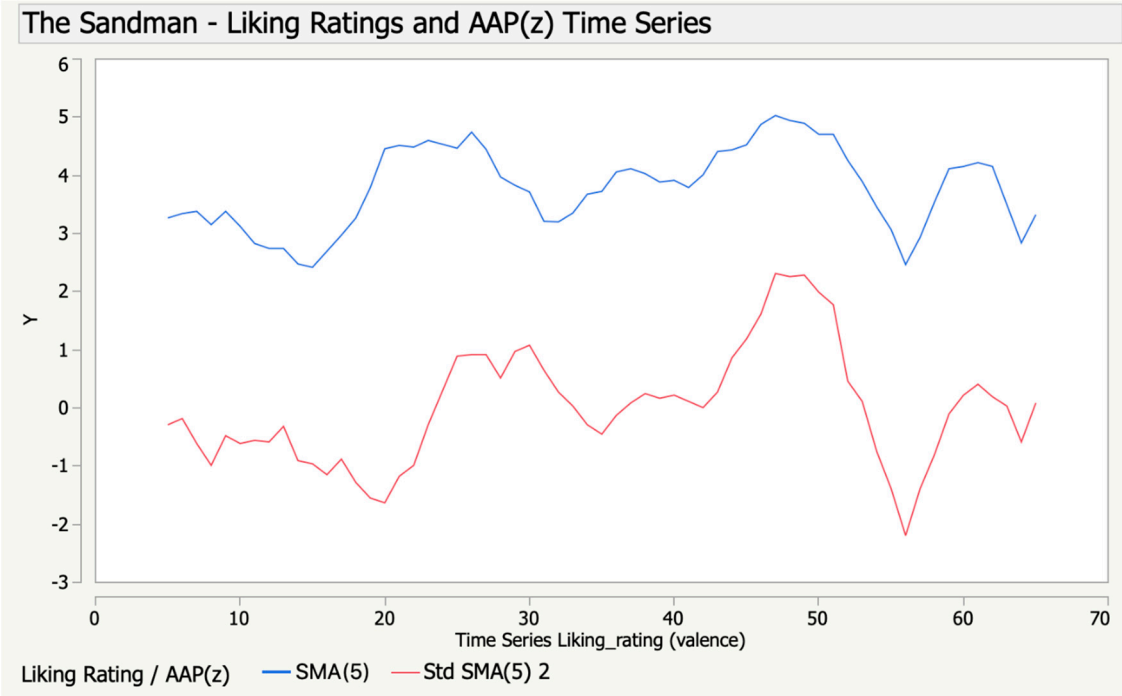


Figure 6. Time series (smoothed average/ smoothed average/SMA, window size = 5) for human liking ratings (in blue) and the corresponding AAP values computed with the winning model C (in red) for 65 text segments from the story *The Sandman*.

Figure 6 shows the smoothed average (SMA, window size = 5) curves for the rating data (in blue) and the corresponding AAP values computed with the winning model C (in red). Actually, the synchrony between the curves is pretty high ($R^2_{adj} = 0.46$, $p < 0.0001$) suggesting that mean AAP for content words is a useful option for predicting the temporal dynamics of human liking ratings. The R^2_{adj} of about 50% sets an upper bound for more sophisticated SA tools that take into account e.g., aspect-based SA [62] or inter- and supralexical text features [12].

3. Summary, Discussion, Limitations and Outlook

The general face validity of three publicly available German VSMs for representing lexico-semantic concepts was established using a tsne approach. The VSMs were then used in the *SentiArt* algorithm to compute the valence and AAP values of the ~120 k words of a German-language database (*subtlex*). In a first cross-validation study, it was shown that the computational *AAP* values predicted ~2800 human valence ratings from the BAWL better than the computational *valence* values, establishing the SDEWAC VSM as the best-fitting of the three VSMs ($R^2 > 0.5$, $r = 0.72$, $p < 0.0001$). A second cross-validation study showed that the *computational* AAP values predicted human liking ratings from an empirical study in which participants read the story *The Sandman* better than *empirically* obtained valence ratings

from the BAWL. It also showed that the time course of human liking ratings was well predicted by the AAP values ($r = 0.65$, $p < 0.0001$).

In sum, the present studies establish *SentiArt*'s AAP variable as a useful predictor of human valence ratings of single words (BAWL) and liking ratings for story segments (*The Sandman*). The predictive validity of the former (R^2 ~0.52) was higher than for the latter (R^2 ~0.23). This could be expected since other than lexical features influence the complex ratings of entire segments or paragraphs. An example is *interlexical* features, which concern the relation between two or more words in a line, sentence, stanza, or paragraph that may well represent dynamic changes or contrasts in readers' affective experience [23,63]. Thus, the *interlexical* features valence and arousal *span* (i.e., the range of valence or arousal values, respectively, of single words across a text segment) are indicative of emotional shifts in a piece of text that can influence readers' mood and indicate an update of the mental situation model [64]. Affective responses to texts can be seen as the dynamic attribution of emotional valence and arousal to every state of the (text) world that an adaptive agent (reader) might visit [23,65]. Valence and arousal spans appear to be appropriate interlexical features serving as proxys for such a dynamic. Indeed, empirical evidence shows that a strong variation in lexical arousal in a piece of text can lead readers on an *emotional rollercoaster* as indicated by online measures of heart-rate variability, brain activity, or liking ratings [57,61,66].

Of course, also supralexical features will affect the emotional experience when reading an entire text. The supralexical features proposed in examples in Jacobs' 4 × 4 matrix for QNAs [23] (*global swing* at the metric level, *global affective meaning* at the phonological level, *syntactic complexity* at the morpho-syntactic level, and *action density* at the semantic level all potentially can affect readers' sentiments and thus be relevant for a future integrative SA tools. However, there is very little research on how these features can best be quantified and integrated into current SA tools [12]. Another important aspect for SAs of entire books concerns the *emotional and figure personality profiles* for main characters as computed by an extended SentiArt algorithm [13]. These profiles can help in predicting the empathy for and identification with story characters, which undoubtedly are an important factor influencing readers' sentiments and moods during the reading of entire novels [67].

Author Contributions: Conceptualization, A.M.J. and A.K.; methodology, A.M.J.; software, A.M.J.; formal analysis, A.M.J. and A.K.; writing, A.M.J. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hogan, P.C. *Affective Narratology: The Emotional Structure of Stories*; The University of Nebraska Press: Lincoln, NE, USA, 2011.
2. Schrott, R.; Jacobs, A.M. *Gehirn und Gedicht: Wie wir Unsere Wirklichkeiten Konstruieren (Brain and Poetry: How we Construct our Realities)*; Hanser: München, Germany, 2011.
3. Jacobs, A.M.; Hofmann, M.J.; Kinder, A. On Elementary Affective Decisions: To Like Or Not to Like, That Is the Question. *Front. Psychol.* **2016**, *7*, 1836. [CrossRef] [PubMed]
4. Lebrecht, S.; Bar, M.; Barrett, L.F.; Tarr, M.J. Micro-valences: Perceiving affective valence in everyday objects. *Front. Psychol.* **2012**, *3*, 107. [CrossRef] [PubMed]
5. Hutto, C.J.; Gilbert, E.E. VADER: A parsimonious rule-based model for sentiment analysis of social media rext. In Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, USA, 4 June 2014.
6. Aryani, A.; Jacobs, A.M.; Conrad, M. Extracting salient sublexical units from written texts: "Emophon," a corpus-based approach to phonological iconicity. *Front. Psychol.* **2013**, *4*, 654. [CrossRef] [PubMed]
7. Aryani, A.; Kraxenberger, M.; Ullrich, S.; Jacobs, A.M.; Conrad, M. Measuring the basic affective tone of poems via phonological saliency and iconicity. *Psychol. Aesthet. Creat. Arts* **2016**, *10*, 191–204. [CrossRef]

8. Bestgen, Y. Can emotional valence in stories be determined from words? *Cognit. Emot.* **1994**, *8*, 21–36. [CrossRef]
9. Dodds, P.S.; Clark, E.M.; Desu, S.; Frank, M.R.; Reagan, A.J.; Williams, J.R.; Mitchell, L.; Harris, K.D.; Kloumann, I.M.; Bagrow, J.P.; et al. Human language reveals a universal positivity bias. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2389–2394. [CrossRef]
10. Jacobs, A.M. Neurocognitive poetics: Methods and models for investigating the neuronal and cognitive-affective bases of literature reception. *Front. Hum. Neurosci.* **2015**, *9*, 186. [CrossRef]
11. Jacobs, A.M. The Gutenberg English Poetry Corpus: Exemplary Quantitative Narrative Analyses. *Front. Digit. Humanit.* **2018**, *5*, 5. [CrossRef]
12. Jacobs, A.M. (Neuro-)Cognitive Poetics and Computational Stylistics. *Sci. Study Lit.* **2018**, *8*, 164–207. [CrossRef]
13. Jacobs, A.M. Sentiment Analysis for Words and Fiction Characters from the Perspective of Computational (Neuro-)Poetics. *Front. Robot. AI* **2019**, *6*, 53. [CrossRef]
14. Jacobs, A.M.; Lüdtke, J. Immersion into narrative and poetic worlds: A neurocognitive poetics perspective. In *Handbook of Narrative Absorption*; Kuijpers, M., Hakemulder, F., Eds.; John Benjamins: Amsterdam, The Netherlands, 2017; pp. 69–96.
15. Jockers, M. Introduction to the Syuzhet Package. Available online: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html (accessed on 15 December 2019).
16. Klinger, R. Digitale Modellierung von Figurenkomplexität am Beispiel des Parzival von Wolfram von Eschenbach. Available online: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/klingern (accessed on 15 December 2019).
17. Klinger, R.; Suliya, S.S.; Reiter, N. Automatic Emotion Detection for Quantitative Literary Studies. A case study based on Franz Kafka's "Das Schloss" and "Amerika". In *Proceedings of the Digital Humanities 2016: Conference Abstracts, 2016*; Jagiellonian University & Pedagogical University: Kraków, Poland, 2016; Available online: http://www.romanklinger.de/publications/klinger-samat-reiter2016.pdf (accessed on 15 December 2019).
18. Kim, E.; Padó, S.; Klinger, R. Investigating the relationship between literary genres and emotional plot development. In Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Vancouver, BC, Canada, 17–26 August 2017.
19. Kim, E.; Padó, S.; Klinger, R. Prototypical emotion developments in adventures, romances, and mystery stories. In *Digital Humanities*; MIT Press: Cambridge, MA, USA, 2017.
20. Kim, E.; Klinger, R. A survey on sentiment and emotion analysis for computation literary studies. *arXiv* **2018**, arXiv:1808.03137v1.
21. Whissell, C. Phonosymbolism and the emotional nature of sounds: Evidence of the preferential use of particular phonemes in texts of differing emotional tone. *Percept. Mot. Skills* **1999**, *89*, 19–48. [CrossRef] [PubMed]
22. Whissell, C.; Fournier, M.; Pelland, R.; Weir, D.; Makarec, K. A dictionary of affect in language: IV. Reliability, validity, and applications. *Percept. Mot. Skills* **1986**, *62*, 875–888. [CrossRef]
23. Jacobs, A.M. Towards a neurocognitive poetics model of literary reading. In *Towards a Cognitive Neuroscience of Natural Language Use*; Willems, R., Ed.; Cambridge University Press: Cambridge, UK, 2015; pp. 135–159.
24. Miller, G.A. *Wörter: Streifzüge durch die Psycholinguistik*; Zweitausendeins: Frankfurt, Germany, 1993.
25. Bühler, K. *Sprachtheorie (Language Theory)*; (Reprint, Stuttgart: Lucius und Lucius); G. Fischer: Stuttgart, Germany, 1934.
26. Freud, S. *Zur Auffassung der Aphasien: Eine kritische Studie [On Aphasia: A Critical Study]*; Deuticke: Wien, Austria, 1891.
27. Jacobs, A.M.; Võ, M.L.-H.; Briesemeister, B.B.; Conrad, M.; Hofmann, M.J.; Kuchinke, L.; Lüdtke, J.; Braun, M. 10 years of BAWLing into affective and aesthetic processes in reading: What are the echoes? *Front. Psychol.* **2015**, *6*, 714. [CrossRef]
28. Ziegler, J.; Montant, M.; Briesemeister, B.; Brink, T.; Wicker, B.; Ponz, A.; Bonnard, M.; Jacobs, A.M.; Braun, M. Do words stink? Neural re-use as a principle for understanding emotions in reading. *J. Cognit. Neurosci.* **2018**, *30*, 1023–1032. [CrossRef]
29. Hu, M.; Liu, B. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; Kim, W., Kohavi, R., Eds.; ACM Press: Washington, DC, USA, 2004; pp. 168–177. [CrossRef]

30. Wiebe, J.; Wilson, T.; Cardie, C. Annotating Expressions of Opinions and Emotions in Language. *Lang. Resour. Eval.* **2005**, *39*, 165–210. [CrossRef]
31. Turney, P.D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the 12th European Conference on Machine Learning, Freiburg, Germany, 5–7 September 2001; Springer: Berlin, Germany, 2001; pp. 491–502.
32. Turney, P.D.; Littman, M.L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst. (TOIS)* **2003**, *21*, 315–346. [CrossRef]
33. Schmidtke, D.S.; Schröder, T.; Jacobs, A.M.; Conrad, M. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behav. Res. Methods* **2014**, *46*, 1108–1118. [CrossRef]
34. Conrad, M.; Recio, G.; Jacobs, A.M. The time course of emotion effects in first and second language processing: Across cultural ERP study with German-Spanish bilinguals. *Front. Psychol.* **2011**, *2*, 1–16. [CrossRef]
35. Hsu, C.-T.; Jacobs, A.M.; Conrad, M. Can Harry Potter still put a spell on us in a second language? An fMRI study on reading emotion-laden literature in late bilinguals. *Cortex* **2015**, *63*, 282–295. [CrossRef]
36. Veltkamp, G.M.; Recio, G.; Jacobs, A.M.; Conrad, M. Is personality modulated by language? *Int. J. Biling.* **2013**, *17*, 496–504. [CrossRef]
37. Hofmann, M.J.; Biemann, C.; Westbury, C.; Murusidze, M.; Conrad, M.; Jacobs, A.M. Simple Co-Occurrence Statistics Reproducibly Predict Association Ratings. *Cognit. Sci.* **2018**, 1–26. [CrossRef] [PubMed]
38. Westbury, C. Pay no attention to that man behind the curtain. *Ment. Lex.* **2016**, *11*, 350–374. [CrossRef]
39. Jacobs, A.M.; Schuster, S.; Xue, S.; Lüdtke, J. What's in the brain that ink may character . . . : A Quantitative Narrative Analysis of Shakespeare's 154 Sonnets for Use in Neurocognitive Poetics. *Sci. Study Lit.* **2017**, *7*, 4–51. [CrossRef]
40. Xue, S.; Lüdtke, J.; Sylvester, T.; Jacobs, A.M. Reading Shakespeare Sonnets: Combining Quantitative Narrative Analysis and Predictive Modeling—An Eye Tracking Study. *J. Eye Mov. Res.* **2019**, *12*. [CrossRef]
41. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
42. Westbury, C.; Keith, J.; Briesemeister, B.B.; Hofmann, M.J.; Jacobs, A.M. Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *Q. J. Exp. Psychol.* **2015**, *68*, 1599–1622. [CrossRef]
43. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. Available online: https://arxiv.org/abs/1301.3781 (accessed on 15 December 2019).
44. Brysbaert, M.; Buchmeier, M.; Conrad, M.; Jacobs, A.M.; Bölte, J.; Böhl, A. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Exp. Psychol.* **2011**, *58*, 412–424. [CrossRef]
45. Van der Maaten, L.J.P.; Hinton, G.E. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2431–2456.
46. Võ, M.L.H.; Jacobs, A.M.; Conrad, M. Cross-validating the Berlin affective word list. *Behav. Res. Methods* **2006**, *38*, 606–609. [CrossRef]
47. Võ, M.L.H.; Conrad, M.; Kuchinke, L.; Hartfeld, K.; Hofmann, M.J.; Jacobs, A.M. The Berlin Affective Word List reloaded (BAWL-R). *Behav. Res. Methods* **2009**, *41*, 534–539. [CrossRef] [PubMed]
48. Baroni, M.; Bernardini, S.; Ferraresi, A.; Zanchetta, E. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Lang. Resour. Eval.* **2009**, *43*, 209–226. [CrossRef]
49. Jacobs, A.M. Quantifying the Beauty of Words: A Neurocognitive Poetics Perspective. *Front. Hum. Neurosci.* **2017**, *11*, 622. [CrossRef] [PubMed]
50. Jacobs, A.M.; Kinder, A. The brain is the prisoner of thought: A machine-learning assisted quantitative narrative analysis of literary metaphors for use in Neurocognitive Poetics. *Metaphor Symb.* **2017**, *32*, 139–160. [CrossRef]
51. Jacobs, A.M.; Kinder, A. What makes a metaphor literary? Answers from two computational studies. *Metaphor Symb.* **2018**, *33*, 85–100. [CrossRef]
52. Jacobs, A.M.; Kinder, A. Features of word similarity. *arXiv* **2018**, arXiv:1808.07999.
53. Marelli, M.; Baroni, M. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychol. Rev.* **2015**, *122*, 485–515. [CrossRef]
54. Ekman, P. Basic emotions. In *Handbook of Cognition and Emotion*; Dalgleish, T., Power, M., Eds.; John Wiley and Sons: Chichester, UK, 1999; pp. 45–60.

55. Warriner, A.B.; Kuperman, V.; Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **2013**, *45*, 1191–1207. [CrossRef]
56. Andrews, M.; Vigliocco, G.; Vinson, D. Integrating experien- tial and distributional data to learn semantic representations. *Psychol. Rev.* **2009**, *116*, 463–498. [CrossRef]
57. Lehne, M.; Engel, P.; Rohrmeier, M.; Menninghaus, W.; Jacobs, A.M.; Koelsch, S. Reading a suspenseful literary text activates brain areas related to social cognition and predictive inference. *PLoS ONE* **2015**, *10*, e0124550. [CrossRef]
58. Remus, R.; Quasthoff, U.; Heyer, G. SentiWS—A Publicly Available German-language Resource for Sentiment Analysis. In Proceedings of the 7th International Language Ressources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010; pp. 1168–1171.
59. Rowling, J.K. *Harry Potter and the Philosopher's Stone*; Bloomsbury: London, UK, 1997.
60. Reagan, A.J.; Mitchell, L.; Kiley, D.; Danforth, C.D.; Dodds, P.S. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* **2016**, *5*, 31. [CrossRef]
61. Wallentin, M.; Nielsen, A.H.; Vuust, P.; Dohn, A.; Roepstorff, A.; Lund, T.E. Amygdala and heart rate variability responses from listening to emotionally intense parts of a story. *Neuroimage* **2011**, *58*, 963–973. [CrossRef] [PubMed]
62. Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*; Cambridge University Press: Cambridge, UK, 2015.
63. Ullrich, S.; Aryani, A.; Kraxenberger, M.; Jacobs, A.M.; Conrad, M. On the relation between the general affective meaning and the basic sublexical, lexical, and interlexical features of poetic texts—A case study using 57 poems of H. M. Enzensberger. *Front. Psychol.* **2017**, *7*, 2073. [CrossRef] [PubMed]
64. Mulcahy, M.; Gouldthorp, B. Positioning the reader: The effect of narrative point-of-view and familiarity of experience on situation model construction. *Lang. Cognit.* **2016**, 96–123. [CrossRef]
65. Joffily, M.; Coricelli, G. Emotional valence and the free-energy principle. *PLoS Comput. Biol.* **2013**, *9*, e1003094. [CrossRef]
66. Hsu, C.T.; Jacobs, A.M.; Citron, F.; Conrad, M. The emotion potential of words and passages in reading Harry Potter—An fMRI study. *Brain Lang.* **2015**, *142*, 96–114. [CrossRef]
67. van Krieken, K.; Hoeken, H.; Sanders, J. Evoking and Measuring Identification with Narrative Characters—A Linguistic Cues Framework. *Front. Psychol.* **2017**, *8*. [CrossRef]