The reader's feeling and text-based emotions: The relationship between subjective self-reports, lexical ratings, and sentiment analysis

Egon Werlen[1]
egon.werlen@ffhs.ch

Fernando Benites[2]
benf@zhaw.ch

Christof Imhof[1]
christof.imhof@ffhs.ch

Per Bergamin[1]
per.bergamin@ffhs.ch

[1]Swiss Distance University of Applied Sciences (FFHS)
[2]Zurich University of Applied Sciences (ZHAW)

Abstract

In this study, we examined how precisely a sentiment analysis and a word list-based lexical analysis predict the emotional valence (as positive or negative emotional states) of 63 emotional short stories. Both the sentiment analysis and the word list-based analysis predicted subjective valence, which however was predicted even more precisely when both analysis methods were combined. These results can, for example, contribute to the development of new technology-based teaching designs, in that positive or negative emotions in the texts or online-contributions of students can be assessed in automated form and transferred into instructional measures. Such instructional actions can, for example, be hints, learning support or feedback adapted to the students' emotional state.

1 Introduction

There has been great progress in technology-based learning in recent decades. Methods and procedures of learning analytics have recently played an important role here. In principle, learning analytics is about collecting data from students during learning and using it to improve teaching. Despite progress in Natural Language Processing (NLP), texts or contributions from students have rarely been used as a source of information for learning analytics or for technology-based learning(e.g. Shibani, 2017). In this article, we used a small corpus of texts with 900 to 1100 characters each in the form of emotional short stories to find out to what extent it is possible to automatically capture emotions as positive or negative emotional colouring of texts. The aim of this article is to assess how well two different methods of automatic capturing of emotions in texts predicted the subjective assessment of emotional reactions to these texts, be it individually or in combination.

2 Theoretical background

In the late nineties, Barrett and Russell (1999) developed the circumplex model, a model of emotions with two dimensions; emotional valence and emotional arousal. Emotional valence is the experience of one's own actual positive or negative feeling. Emotional arousal is the subjective amount of internal activation or energy. Together, these two dimensions form the core affect, "the most elementary consciously accessible affective feelings that need not be directed at anything" (S. 806). The circumplex model provided the theoretical basis for the present work.

Emotional valence, based on the circumplex model, was measured on a bipolar scale, ranging form very negative to very positive. This method was originally conceived by Wundt (1896) and is the most commonly used method to date. However, like the sentiment analysis used in this study, some theories view valence as a bivariate construct (e.g. Norris et al., 2010; Briesemeister et al., 2012; Shuman et al., 2013; Kron et al., 2015). According to those views, humans can perceive objects (e.g. images, words, texts) as positive and negative at the same time, enabling them to have an ambiguous quality. This highlights that emotion measurements are a challenging and debated task (see also e.g. Mauss and Robinson, 2009).

2.1 Subjective measurement by self-reporting

Today, research assumes that individual measurements cannot capture the phenomenon of emotions entirely. This leads to the practice of using multiple measuring methods in scientific investigations, often in conjunction. Self-reports such as questionnaires or single item questions are a popular way of measuring emotions, and for good reasons: they have good validity (as long as response biases are

taken into account) and enable quick and simple data collection. Non-verbal alternatives to measure emotions can also be used, such as the Self-Assessment Manikins (SAM scale) from Bradley and Lang (1994), measuring feelings, i.e. the subjective experience of emotions. This instrument contains visual rather than verbal stimuli (i.e. pictures rather than questions), which consist of abstract representations of a human being displaying different emotions. The scale varies in three dimensions; valence, arousal, and dominance. The valence dimension shows pictures ranging from a smiling face to a frowning face, with more neutral expressions in-between; and in the arousal dimension, pictures range from a sleepy and calm figure to a wide-eyed, excited expression. We did not use the dominance dimension that represents the controlling and dominant nature of emotion shown by a tiny figure in the middle of a square for low dominance towards a oversize figure going beyond the borders of the square for high dominance. Raters were instructed to choose the image that best represents their own current emotional state.

2.2 Objective measurement by lexical ratings, and sentiment analysis

Despite their popularity, self-reports are far from the only instrument being used in affective science. Lexical analysis (i.e. analysis based on single words) is a different, more objective instrument which historically has been used significantly less often. In this regard, Jacobs et al. (2015) argues, based on long existing works by Freud (1891) and Bühler (1934), that spoken or written words contain the potential to elicit both overt or covert sensu-motoric or affective reactions. In this context we speak of embodied stimuli. Recent neurological research supports this relationship as demonstrated in Jacobs (2015). On the basis of these, it can be explained that words can evoke both basic and fictional emotions as well as something like aesthetic feelings.

Before neurological research pointed out these connections, there was a clear language-emotion gap, i.e. most emotion theories ignored language functions, while linguistic theories ignored affective processes. In order to bridge that gap, the Berlin Affective Word List (BAWL-R) was developed (Vo et al., 2009). The BAWL-R is a large German word list containing almost 3000 words (nouns, verbs, and adjectives) from the CELEX database (Baayen et al., 1993), each rated on valence, arousal, and imageability. The list also includes psycholinguistic factors (e.g. number of letters, phonemes, word frequency, accent). It is free for download ([1]). To open the file a password must be requested. The BAWL-R enables estimations of the emotional potential for single words but also extrapolations for sentences and whole texts.

The BAWL-R specifically has been utilized for this purpose as well: Aryani et al. (2015) analysed poems, Lehne et al. (2015) examined E.T.A. Hoffmann's black-romantic story "The Sandman", Hsu et al. (2015) analysed passages of Harry Potter novels, and Jacobs and Kinder (2017) inquired potentially relevant properties of Skakespear's sonnets. These studies found out that affective word ratings correlated with whole text ratings and came to the conclusion that a text's constituting words can predict its emotional potential. Studies using the BAWL-R to predict subjective valence of short texts (Hsu et al., 2015) and poems (Ullrich et al., 2017) with lexical valence found correlations of $r = .58$ (short texts) and $r = .65$ (poems).

Since about the year 2000, research on sentiment and, as a consequence thereof, the term *sentiment analysis* appeared in scientific literature of computational science with increased frequency e.g. (Nasukawa and Yi, 2003; Das and Chan, 2001). Liu (2012) describes sentiment analysis as part of natural language processing (NLP) that extracts people's emotions, sentiments, opinions etc. out of spoken or written language. It focuses mainly on positive and negative sentiments. Sentiment analysis is a learning-based approach, that - in contrast to lexical analysis - does not necessarily rely on rated word lists and instead implements machine learning. Technically speaking, word-based lexical analysis could be categorized as a semantic approach to sentiment analysis that does not necessarily implement machine learning. Sentiment analysis, also called opinion mining or polarity detection, as explained by Fueyo (2018), "refers to the set of AI algorithms and techniques used to extract the polarity of a given document: whether the document is positive, negative or neutral" that is represented as classes or a probability. Angiani et al. (2016) lists possible steps of a sentiment analysis: 1) initialization step (data collection, data processing, attribute selection), 2) learning step (algorithm, training model), and 3) evaluation step (test set).

The automatic sentiment analysis system used for this paper is composed of two parts, namely the model and the data. The multi-layered convolutional network model is the same as in Deriu et al. (2017). The authors trained this network as shown in Figure 1 with a large number of tweets in different languages that were weakly supervised, and demonstrated the importance of using pre-training of such networks. The specific pre-training procedure, named distant-supervised learning, is trained

[1] cf. https://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/allgpsy/Download/BAWL/index.html accessed May 2019

on larger weakly or non-labelled samples[2]. Afterwards the network is further trained on a much smaller data set with manually strongly labelled samples. The approach was evaluated on various multi-lingual data sets, including the SemEval-2016 sentiment prediction benchmark (Task 4), where it achieved state-of-the-art performance.
This model was trained on the SB10k German Twitter sentiment corpus (Cieliebak et al., 2017), which is a corpus for sentiment analysis with approximately 10,000 German tweets. Tweets are normally a sentence long and are often connoted with emotions. Although the domain is not the same, the focus on sentence and on emotions is very similar in the used data sets (train and test). The used word embeddings were weakly trained on 40 millions German tweets. Here, emoticons were used for automatically labelling the emotional content of a tweet (positive, negative, neutral). Finally, the output of the network is the confidence (from 0 to 1) for each one of the three sentiments.
Both lexical and sentimental analyses have been applied to different types of texts to measure their emotional potential in different contexts. Mossholder et al. (1995) analysed emotions in open-ended survey responses by applying the Dictionary of Affect in Language (DAL); Loughran and McDonald (2015) used the Diction software in order to analyse and categorize the tone of business documents such as financial reports; Humphreys and Wang (2017) implemented automated text analysis for examining text patterns in consumer research; Lima et al. (2015) analysed Twitter messages within a polarity analysis framework, Whissell (2011) analysed Poe's poetry and Whissell (1996) used the "emotion clock" to conduct a stylometric analysis of Beatles songs, to name a few examples.

2.3 Combination of different measurement procedures

A comparison of different procedures for emotion recognition on the sentence level was conducted by Aman (2007). He concluded that a combination of different automatic procedures for recording emotions is advantageous. This finding is also supported in a paper by Strapparava and Mihalcea (2010). They tested several methods for automatically detecting emotions in short texts (headlines and blog posts; 100-400 characters). Six headline advisors rated the presence of six distinct emotions as well as the valence of the texts, which were then predicted by several procedures. The study found that "different methods have different strengths, especially with respect to individual emotions" (p. 35). Most interestingly, the correlation between emotions evaluated by human raters and those found by algorithms was moderate with max. $r = .48$ (explanation of variance max. 24%). The largest effect was found with valence analysed in a knowledge-based, area-independent, unsupervised CLaC approach. We assume, as already mentioned above, that different measurements can cover certain aspects of the complex phenomenon emotions that others do not. Different measurements often only reveal parts of a phenomena and might sometimes even be contradictory. Thus, the combination of several measurement techniques can prove to be fruitful.
In our study, we are interested in the combination between lexical analysis, sentiment analysis and self-report and specifically, if prediction of the latter improves when the former two are combined.

2.4 Hypotheses

As discussed above, it has been known for a long time (e.g. Freud (1891); Bühler (1934) that words can trigger emotional reactions, which more recently has been confirmed in neurological studies Jacobs (2015). According to the circumplex model of emotion by Barrett and Russell (1999), the emotional valence, i.e. the personal appraisal whether and how strongly something is perceived positively or negatively, is one of the most basic emotional reactions. The emotional valence, i.e. the subjective valence, of the 63 short texts was assessed by university students rating their emotional responses to these texts (17-19 ratings per text). As explained above, the emotional valence of a text can also be measured objectively, in our case with sentiment analysis and lexical analysis. We were interested in finding out if these automated objective measurement approaches could predict the subjective valence. If so, they could serve as an approximation rather than relying on repeated self-reports of subjective ratings. This leads to the first hypothesis.
As several studies have shown (e.g. Aman (2007); Strapparava and Mihalcea (2010), combinations of several methods for estimating emotions in texts lead to better predictions than one method alone. This leads to the second hypothesis.

1. The emotional valence measured by lexical analysis and by sentiment analysis each predict the subjective valence of the short texts.

2. The combination of the measurements methods (lexical analysis and sentiment analysis) increases the predictive power.

Despite the sizable amount of research in emotion and in text analysis, we are not aware of many studies that not only compared (e.g. Nielsen, 2011; Hutto and Gilbert, 2014) but also combined both word-list-based lexical analysis and

[2]On twitter emoticons/emojis can be used as weak labels, for instance a tweet with a smiling emoji will probably have a positive sentiment.

sentiment analysis to predict subjective ratings of emotional valence in short texts (e.g. Dhaoui et al., 2017).

3 Methods

3.1 Samples and measurements

The 63 analysed texts originated from a collection of 102 German texts written by 32 authors, 21 of which were German speaking students and staff of an University in Germany (mean age 26.10, SD = 10.65; gender: 85% women), and 11 of which were recruited by the first author (university staff and people recruited via social media and personal contacts; mean age 36.82, SD = 15.78; 64% women). These 63 texts are part of an international database with over 200 emotional short stories which are developed and refined within the framework of the COST initiative E-Read IS 1404 (Kaakinen et al., in preparation). The international database contains stories from Finland, France, Germany, Portugal, Spain, Switzerland, and Turkey. All stories are subjectively rated on emotional valence, emotional arousal and comprehensibility in their original language and in English. All texts have a length of 900 to 1100 characters including spaces. Texts that were not written in the first person were rewritten without changing their content and structure. The topic varies from story to story, some of them tell of joyful events and experiences (e.g. birth, love, music) or negative ones (e.g. death, abuse). A few stories are emotionally neutral, i.e. neither positive nor negative and with a medium level of emotional arousal. The stories are mostly easy to understand. Once finished, the database will be presented in a publication and made freely accessible.

The subjective valence rating of the texts was conducted with the Self-Assessment-Manikin scale (SAM[3]) by Lang (1980). We used a modified 9-point scale by Suk (2006). Participants were instructed to rate the texts by choosing one of nine icons to represent their current emotional state. The 63 texts were rated on the survey platform Qualtrics by 55 native German speaking university students from different majors of a German University. The raters' mean age was 23.47 years (SD = 2.62), 90.9% were female. Each participant rated a randomly predetermined set of 21 texts in randomized order, so that each text was evaluated by one of three groups with each 17-19 participants. As compensation, participants had the chance to win one of fifteen 10 € Amazon vouchers. The inter-rater reliability for the subjective rating of emotional valence calculated with the R package irr by Gamer et al. (2012) was .98 or more in each of the three groups. The semantic lexical analysis of the text was conducted with the revised form of the Berlin Affective Word List BAWL-R (Vo et al., 2009) in R (R Core Team, 2017), using the packages tidyverse (Wickham, 2017) and sylly (Michalke, 2018). In that list, valence had been rated on a 7-point Likert scale (-3 very negative through 0 neutral to +3 very positive). For each short story, we averaged the valence of all the words in that text represented in the BAWL-R.

The automatic sentiment analysis was trained on sentences. Nevertheless, we applied it to our short stories as a whole instead, since the subjective ratings we wanted to predict were on a text rather than a sentence level. For this paper, we calculated a new overall valence variable for the sentiment analysis data based on the negative and positive scores (negative sentiment minus positive sentiment), assuming that the neutral sentiment had no influence on the positive or negative orientation of the analysis. The reason for this decision was that the three original variables sum up to 1 and are therefore interdependent. Consequently, their individual effects on the subjective ratings canceled each other out. In order to obtain values comparable to the BAWL-R valence variable, this new variable was created. We further analyzed the text in terms of readability. Readability was scored with the well established Flesch Index (Flesch, 1948), using a formula adapted to the German language (Amstad, 1978). Means and standard deviations of the all used measures for valence are reported in Table 1.

Valence	mean	SD	min	max	scale
Subjective	4.46	2.21	1.17	8.61	-3 - 3
Lexical	0.63	0.27	0.02	1.17	1 - 9
Sentiment	-0.44	0.26	-0.95	0.20	-1 - 1
Sentiment					
positive	0.14	0.10	0.01	0.47	0 - 1
negative	0.29	0.12	0.04	0.84	0 - 1

Table 1: Mean, standard deviation (SD), minimum (min), maximum (max), and possible values (scale) of valence measured subjective (rated by students), with lexical analysis (BAWL-R), and with sentiment analysis

3.2 Analyses

We chose to conduct our regression analyses with a Bayesian approach, which has important advantages over the traditional frequentist null hypothesis significance testing. Within the

[3] cf. http://irtel.uni-mannheim.de/pxlab/demos/index_SAM.html accessed Feb. 2019

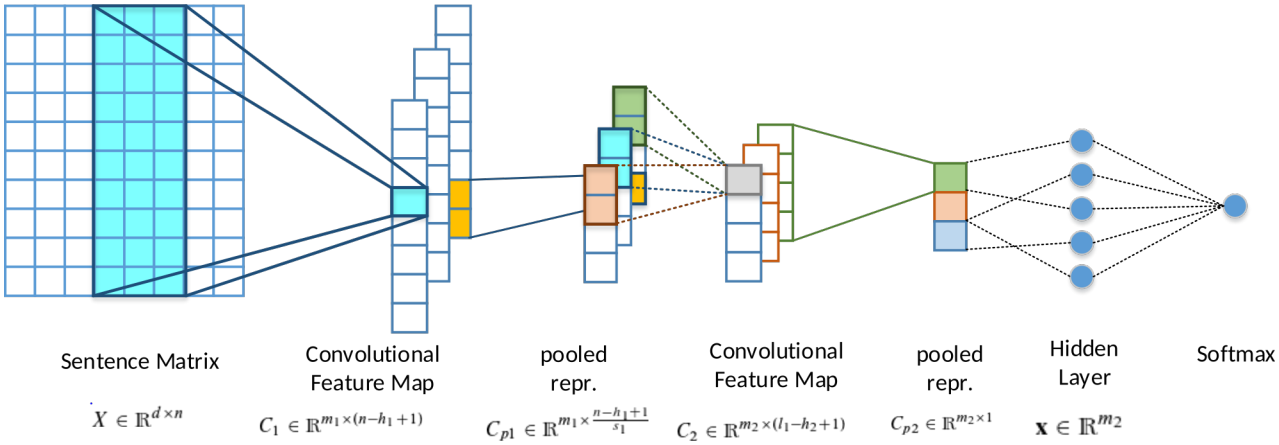

Figure 1: Convolutional Neural Network Model from Deriu et al. (2017)

Bayesian approach, the interpretation of data is not affected by sampling intention. In contrast to the frequentist approach, the Bayesian approach permits assessment of the relative credibility of parameter values given the data and the statistical model (Kruschke, 2010). The statistical analyses were conducted with R version 3.3.4 (R Core Team, 2017) and the R package *brms* version 2.4.0 (Bürkner, 2018), which is a package for Bayesian generalized multivariate non-linear multilevel models. To allow comparisons with other studies that correlated lexical or sentimental analysis with subjective ratings of texts, we calculated correlations of the standardized values averaged over the 63 texts as beta values with brms. For the multilevel models predicting subjective valence, the raw data of all 55 raters were included in the model with rater as a level 2 predictor. The resulting sample included 1143 observations, i.e. 63 texts with an average of 18 raters. The predictors (sentiment, lexical valence, and Flesch Index) were averaged for each of the 63 texts. The subjective valence ratings - an ordinal scaled variable with values ranging from 1 to 9 - were modelled with a cumulative distribution. The Bayesian Credible Interval, meaning the range a certain value lies within with a probability of 95% (thus not to be confused with the frequentist Confidence Interval!) is reported for all results. Since this is the first study in this context applying Bayesian analysis, no informative priors were available. We thus decided to use brms' default priors. The Leave-One-Out Cross-Validation information criteria (LOOic) was used to compare the different models. The LOOic is a method "for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values" (p. 1413; (Vehtari et al., 2017)).

4 Results

The correlation between the sentiment value - calculated as the difference between negative and positive sentiment - and the lexical value of valence was $r = .50$ (95% Credible Interval - CrI = [.28; .72]). Both of them had a moderate positive correlation with the subjective valence ratings of the texts ($r = .51$ (95% CrI = [.28; .72] for sentiment; $r = .62$ (95% CrI = [.42; .82] for lexical valence). There was a weak correlation between the Flesch readability score and the other three variables (sentiment: $r = -.24$ (95% CrI = [-.48; .01]; lexical valence: $r = -.12$ 95% CrI = [-.37; .12]; subjective valence: $r = -.24$ (95% CrI = [-.50; .01]).
A visual inspection of the MCMC chains and the R-hat diagnostic with all R-hat values < 1.02 revealed good convergence for all estimated parameters of all calculated models.
The restricted model (Model 0) including the intercepts and the level 2 variable only had a LOOic of 4969. Model 1 predicting subjective valence by sentiment had an effect of $\beta = 4.60$ (95% CrI = [2.62; 6.56]). The LOOic was 4717. Model 2 predicting subjective valence by lexical valence had an effect of $\beta = 5.37$ (95% CrI = [3.62; 7.01]) with a LOOic of 4578. To decide, which model is to prefer, we relied on the credible intervals of the LOOic. The credible intervals of the LOOic of Model 1 and 2 did not overlap with the LOOic of the restricted model 0 (see Table 2). That lead to our conclusion that both models predicted subjective valence and that the first hypothesis could be confirmed.

Model 3 predicting subjective valence of texts by sentiment and lexical valence (BAWL-R) is presented in Table 3. The design formula for model 3 was formulated as follows:

Model	LOOic	se	CrI 5%	CrI 95%
$M0$	4969	17	4935	5002
$M1$	4717	36	4646	4787
$M2$	4578	39	4502	4654
$M3$	4515	42	4433	4597
$M3+$	4494	42	4412	4577

Table 2: LOO information criteria with standard error and credible intervals (CrI)

$$R_i \sim \text{Ordered}(\mathbf{p}) \quad \text{[likelihood]}$$
$$\text{logit}(p_k) = \alpha_k - \phi_i \quad \text{[cumulative link and linear model]}$$
$$\phi_i = \beta_{BAWL} BAWL_i + \beta_{Sent} Sent_i \quad \text{[linear model]}$$
$$\alpha_k \sim \text{Normal}(0, 10) \quad \text{[common prior for each intercept]}$$
$$\beta_{BAWL} \sim \text{Normal}(0, 10) \quad [\beta_{BAWL} \text{ prior}]$$
$$\beta_{Sent} \sim \text{Normal}(0, 10) \quad [\beta_{Sent} \text{ prior}]$$

R_i is the ordered distribution (i.e. a categorical distribution that takes the vector $\mathbf{p} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8\}$ of probabilities of each subjective valence rating value below the maximum category of 9). α_k is the unique intercept of each possible outcome value k, ϕ_i is the linear model that is subtracted from each intercept, β_{BAWL} and β_{Sent} are the slopes of the BAWL-R (lexical analysis) and sentiment values respectively and $BAWL_i$ and $Sent_i$ are the values of both predictor variables on row i.

Parameter	$\hat{R}$	n_eff	β	CrI 5%	95%
$b_{raterIntercept}$	1.01	1242	0.11	0.01	0.28
$b_{sentiment}$	1.00	4000	2.07	1.58	2.59
$b_{valence}$	1.00	4000	3.42	2.96	3.89

Table 3: Results of Bayesian linear regression analysis

The subjective valence was predicted by sentiment with $\beta = 2.07$ (95% CrI = [1.58; 2.59]), and by lexical valence with $\beta = 3.42$ (95% CrI = [2.96; 3.89]). The LOO information criteria of model 3 (LOOic = 4515) was smaller than that of either of the other models. The credible interval of model 1, but not of model 2 does not overlap with the credible interval of the combined model 3. We conclude that sentiment and lexical analysis predict subjective valence better than sentiment analysis alone. Even if the credible interval of model 3 overlaps with the credible interval of model 2, we consider the difference of their LOOic big enough to conclude that model 3, i.e. the prediction of subjective valence is better when sentiment and lexical analysis are combined than either one of them on their one. This confirmed the second hypothesis. The level 2 predictor (text raters) in model 3 had a small but negligible effect on subjective valence.

The integration of readability (Flesch-Index) did not improve the model. The credible intervals were mostly overlapping and the information gain was negligible with a small effect of readability on subjective valence ($\beta = -0.04$; 95% CrI = [-0.09; -0.02]; LOOic = 4494). The LOOic, standard errors, and credible intervals of the five models are listed in Table 2.

In Figure 2, the prediction of subjective valence by sentiment and lexical valence given the data and the statistical model considering the effects of both predictors (model 3) is visualized. The figure shows the slope (blue line) with its 95% gray shadowed credible interval. Both predictors have a tight credible interval that does not include 0, indicating clear positive effects for both predictors.

5 Discussion

The aim of this study was to compare different techniques to capture emotions in 63 short texts. Our investigations focused on the question whether the prediction of readers' subjectively appraised emotions towards texts improves when word list-based lexical analysis and sentiment analysis are both considered. The results confirmed our hypotheses that lexical and sentiment analyses both predict subjective valence independently or in combination (hypothesis 1). The strongest effect resulted when both approaches were combined (hypothesis 2). That confirms Dhaoui et al. (2017) and corresponds to findings of Aman (2007) and Strapparava and Mihalcea (2010) that combinations of algorithms result in better predictions.

Other studies predicting subjective valence ratings of texts with the same word list (BAWL-R) found correlations of $r = .58$ (Hsu et al. (2015); for short passages of the Harry Potter novels) and $r = .65$ ((Ullrich et al., 2017); for poems of Enzensberger), which are in the same range as our results ($r = .62$). There are studies for instance Settanni and Marengo (2015) using other word list (e.g. LIWC) with lower correlations between negative emotions expressed on Facebook posts and corresponding subjective negative emotions ($r = .22$, for younger people $r = .40$). Similar is found for the sentiment analysis, where our results ($r = .50$) correspond to results published in the literature. Correlations of different algorithms with subjective valence ratings were reported for instance by

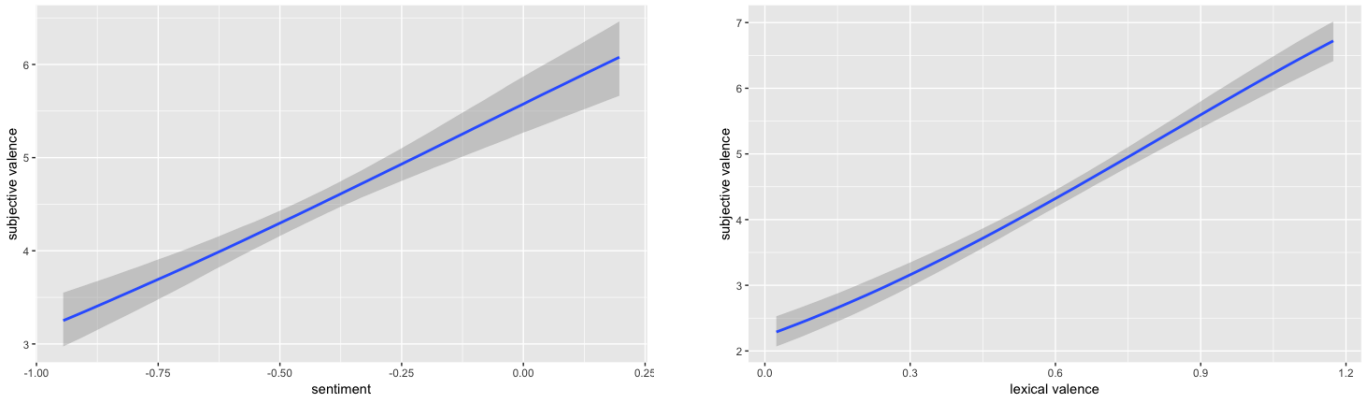

Figure 2: Prediction of subjective valence by sentiment and lexical valence

Strapparava and Mihalcea (2010) in detecting sentiment in headlines. The algorithm with the best predictive power was the CLaC system that "relies on a knowledge-based domain-independent unsupervised approach to headline valence detection and scoring. The system uses three main kinds of knowledge: a list of sentiment-bearing words, a list of valence shifters and a set of rules that define the scope and the result of the combination of sentiment-bearing words and valence shifters" (p. 28). This algorithm found a correlation of $r = .48$ for valence. The correlations with the other four algorithms were all below $r = .40$. In comparison to other studies, the sentiment analysis used in this study revealed a rather high correlation. A more recent study (Preoţiuc-Pietro et al., 2016) found a higher correlation of $r = .65$ between sentiment analysis and subjective valence with a bag-of-words linear regression model.

When both measurement techniques were combined in a model, the effect of lexical valence predicting subjective valence was stronger than the effect of the sentiment analysis. The $\beta = 2.07$ in model 3 for sentiment means that an increase of 1 SD in the sentiment values corresponds to an increase of 2.07 SDs in the predicted subjective valence. Likewise, the $\beta = 3.42$ for lexical valence means that an increase of 1 SD in the lexical valence values corresponds to an increase of 3.42 SDs in the predicted subjective valence. This stronger effect of lexical analysis was also visible in the correlations of each variables with subjective valence. Both predictors correlate with each other ($r = .50$), and therefore share a good part of their variance. This explains that overlap between the credible intervals of model 3 (sentiment and lexical analysis as predictors) and model 2 (lexical analysis as predictors). Nevertheless, we considered the information gain of model 3 over model 2 to be large enough and therefore favour model 3.

It is known from literature that the difficulty of texts has an impact on the emotions when reading (e.g. Yin et al., 2014; Ben-David et al., 2016). To take this into account we investigated an additional model 3+. One way to determine the text difficulty is with the Flesh-Index (Flesch, 1948; Amstad, 1978). When this predictor was taken into account in the model, only a very small effect could be found. Therefore, we decided not to pursue this additional variant any further. One reason for the small effect of the Flesh-Index might be that when selecting the texts at the beginning of the study we made sure that none of the texts used had extreme Flesch values in order to avoid biases of the measurement results due to comprehension problems.

An explanation for the weaker performance of the sentiment analysis compared to the lexical rating may be that the 63 analysed texts were part of an international database with emotional short texts (Kaakinen et al., in preparation). In this context, the emotional content of the entire text (not on the word or sentence level) was assessed by student raters. This differs from the method of sentiment analysis, which was applied on each short story but was trained on Twitter messages, i.e. sentence level (Cieliebak et al., 2017). As in other studies, the lexical analysis is based on the average valence of words that previously were evaluated by students (Vo et al., 2009). We assigned each word of the short texts, which was also included in the Berlin Affective Word List, its valence value and averaged these values getting a mean value for each short text. We assume that due to different aspects of valence measured in the procedures mentioned, the combination model and therefore the combination of the different aspects of valence measured achieved the best prediction values. However, in order to actually confirm this assumption we need to further investigate whether the correlations between the three measurements and the fit of the different models remain at a lower taxonomic level, i.e. at the sentence respectively word level instead

of the text level, and observe the predictive power accordingly. Another question is whether the combination of measurement techniques developed and validated in a context other than short stories, such as the sentiment analysis using tweets, is appropriate, or whether it is better to use other techniques developed in the same context. We are under the impression, that the sentiment analysis applied in this study did a pretty good job compared to other procedures.

There are some aspects to our approach that we did not account for in this study that may be worth exploring in future studies. One such aspect is the perceived difficulty of the rating task since the subjective ratings may be biased if the task is thought to be either particularly easy or particularly difficult. This concerns both the text ratings as well as the ratings that resulted in the two analysis approaches that rely on subjective expert ratings at their very core. Another aspect worthy of inspection is the discrepancy between human ratings and both analysis approaches since they do not necessarily align at all times. Exploring under which circumstances they diverge may prove to be a promising venture.

6 Conclusions

The results indicate that lexical and sentiment analyses predict subjective appraisal of emotions triggered by short texts. The two methods are not redundant. It is therefore worthwhile analyzing the emotional potential of texts applying both measurement procedures. A next step is to repeat these analyses on sentence and on word level to check whether we get an even stronger predictive power. We also need to examine the integration of other text properties, including other semantic parameters, into our analysis, as done by Jacobs and Kinder (2017). The small effect gain of the Flesch-Index can be interpreted as an indication that non-emotional text properties could play a role in the perception of emotions in a text.

The results found, for example, can contribute to the development of new instructional designs that measure emotional appraisals of students engaged in digital learning tasks. Positive or negative emotions in the texts or online-contributions of students can be assessed in automated form and transferred into instructional measures, and thus help to integrate automated learning support into feedback, hints or adaptive instructional design. We need even more predictive power for useful integration of such sensors, i.e. measurements of emotional or affective properties of texts in digital learning, in educational practice. From our point of view, this can be achieved by combining different measurement methods

Acknowledgments

A big 'Thank you' to Yvonne Kammerer of the Leibnitz-Institut für Wissensmedien in Tübingen. She organized the collecting of the text from Germany, and the rating of the 63 texts that were collected as part of a project in the COST Action E-READ. We also thank Mark Cieliebak and Jan Deriu for providing the sentiment prediction system and the helpful discussions, and Stéphanie McGarrity for proofreading and useful suggestions.

References

Saima Aman. 2007. *Recognizing emotions in text*. Ph.D. thesis, University of Ottawa (Canada).

T Amstad. 1978. Wie verständlich sind unsere zeitungen?[how understandable are our newspapers?]. *Unpublished doctoral dissertation, University of Zürich, Switzerland* .

Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi. 2016. A comparison between preprocessing techniques for sentiment analysis in twitter. In *KDWeb*.

A Aryani, M Kraxenberger, S Ullrich, AM Jacobs, and M Conrad. 2015. Measuring the basic affective tone in poetry using phonological iconicity and subsyllabic salience. *Psychol. Aesthet. Creat. Arts* .

R Harald Baayen, Richard Piepenbrock, and H Van Rijn. 1993. The celex lexical database (cd-rom). linguistic data consortium. *Philadelphia, PA: University of Pennsylvania* .

Lisa Feldman Barrett and James A Russell. 1999. The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science* 8(1):10–14.

Boaz M Ben-David, Maroof I Moral, Aravind K Namasivayam, Hadas Erel, and Pascal HHM van Lieshout. 2016. Linguistic and emotional-valence characteristics of reading passages for clinical use and research. *Journal of fluency disorders* 49:1–12.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry* 25(1):49–59.

Benny B Briesemeister, Lars Kuchinke, and Arthur M Jacobs. 2012. Emotional valence: A bipolar continuum or two independent dimensions? *SAGE Open* 2(4):2158244012466558.

Karl Bühler. 1934. Sprachtheorie (language theory). *Stuttgart: G. Fischer* .

Paul-Christian Bürkner. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1):395–411.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *5th International Workshop on Natural Language Processing for Social Media, Boston, MA, USA, December 11, 2017*. Association for Computational Linguistics, pages 45–51.

Sanjiv Das and M Chan. 2001. Extracting market sentiment from stock message boards. *Asia Pacific Finance Association* 2001.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In *WWW 2017 - International World Wide Web Conference*. Perth, Australia.

Chedia Dhaoui, Cynthia M Webster, and Lay Peng Tan. 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing* 34(6):480–488.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32(3):221.

Sigmund Freud. 1891. *Zur auffassung der aphasien: eine kritische studie*. F. Deuticke.

Enrique Fueyo. 2018. Understanding what is behind sentiment analysis (part i). Last accessed 6 March 2019. https://building.lang.ai/understanding-what-is-behind-sentiment-analysis-part-i-eaf1bcb43d2d.

Matthias Gamer, Jim Lemon, Maintainer Matthias Gamer, A Robinson, and W Kendall's. 2012. Package 'irr'. *Various coefficients of interrater reliability and agreement* .

Chun-Ting Hsu, Arthur M Jacobs, Francesca MM Citron, and Markus Conrad. 2015. The emotion potential of words and passages in reading harry potter–an fmri study. *Brain and language* 142:96–114.

Ashlee Humphreys and Rebecca Jen-Hui Wang. 2017. Automated text analysis for consumer research. *Journal of Consumer Research* 44(6):1274–1306.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Arthur M Jacobs. 2015. Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception. *Frontiers in human neuroscience* 9:186.

Arthur M Jacobs and Annette Kinder. 2017. "the brain is the prisoner of thought": A machine-learning assisted quantitative narrative analysis of literary metaphors for use in neurocognitive poetics. *Metaphor and Symbol* 32(3):139–160.

Arthur M Jacobs, Melissa L-H Võ, Benny B Briesemeister, Markus Conrad, Markus J Hofmann, Lars Kuchinke, Jana Lüdtke, and Mario Braun. 2015. 10 years of bawling into affective and aesthetic processes in reading: what are the echoes? *Frontiers in psychology* 6:714.

Johanna K Kaakinen, Egon Werlen, Yvonne Kammerer, S Ruiz-Fernandez, Cengiz Acartürk, Xavier Aparicio, Thierry Baccino, Ugo Ballenghein, Per Bergamin, Nuria Castells Gomez, Armanda Costa, Isabel Falé, Olga Megalakaki, and M Minguela. in preparation. Emotional text database [working title]. *Behavior research methods* .

Assaf Kron, Maryna Pilkiw, Jasmin Banaei, Ariel Goldstein, and Adam Keith Anderson. 2015. Are valence and arousal separable in emotional experience? *Emotion* 15(1):35.

John K Kruschke. 2010. What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences* 14(7):293–300.

PJ Lang. 1980. Self-assessment manikin. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida* .

Moritz Lehne, Philipp Engel, Martin Rohrmeier, Winfried Menninghaus, Arthur M Jacobs, and Stefan Koelsch. 2015. Reading a suspenseful literary text activates brain areas related to social cognition and predictive inference. *PLoS One* 10(5):e0124550.

Ana Carolina ES Lima, Leandro Nunes de Castro, and Juan M Corchado. 2015. A polarity analysis framework for twitter messages. *Applied Mathematics and Computation* 270:756–767.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Tim Loughran and Bill McDonald. 2015. The use of word lists in textual analysis. *Journal of Behavioral Finance* 16(1):1–11.

Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition and emotion* 23(2):209–237.

Meik Michalke. 2018. *sylly: Hyphenation and Syllable Counting for Text Analysis*. (Version 0.1-5). https://reaktanz.de/?c=hacking&s=sylly.

Kevin W Mossholder, Randall P Settoon, Stanley G Harris, and Achilles A Armenakis. 1995. Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of management* 21(2):335–355.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*. ACM, pages 70–77.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* .

Catherine J Norris, Jackie Gollan, Gary G Berntson, and John T Cacioppo. 2010. The current status of research on the structure of evaluative space. *Biological psychology* 84(3):422–436.

Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 9–15.

R Core Team. 2017. A language and environment for statistical computing http://www. r-project. org.

Michele Settanni and Davide Marengo. 2015. Sharing feelings online: studying emotional well-being via automated text analysis of facebook posts. *Frontiers in psychology* 6:1045.

Antonette Shibani. 2017. Combining automated and peer feedback for effective learning design in writing practices. In *ICCE 2017-25th International Conference on Computers in Education: Technology and Innovation: Computer-Based Educational Systems for the 21st Century, Doctoral Student Consortia Proceedings*.

Vera Shuman, David Sander, and Klaus R Scherer. 2013. Levels of valence. *Frontiers in Psychology* 4:261.

Carlo Strapparava and Rada Mihalcea. 2010. Annotating and identifying emotions in text. In *Intelligent Information Access*, Springer, pages 21–38.

Hyeon-Jeong Suk. 2006. *Color and Emotion-a study on the affective judgment across media and in relation to visual stimuli*. Ph.D. thesis, Universität Mannheim.

Susann Ullrich, Arash Aryani, Maria Kraxenberger, Arthur M Jacobs, and Markus Conrad. 2017. On the relation between the general affective meaning and the basic sublexical, lexical, and inter-lexical features of poetic texts—a case study using 57 poems of hm enzensberger. *Frontiers in psychology* 7:2073.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing* 27(5):1413–1432.

Melissa LH Vo, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods* 41(2):534–538.

Cynthia Whissell. 1996. Traditional and emotional stylometric analysis of the songs of beatles paul mccartney and john lennon. *Computers and the Humanities* 30(3):257–265.

Cynthia Whissell. 2011. To those who feel rather than to those who think: Sound and emotion in poes poetry. *International Journal of English and Literature* 2(6):149–156.

Hadley Wickham. 2017. The tidyverse. *R package ver. 1.1. 1* .

Wilhelm Wundt. 1896. *Grundriss der Psychologie*. Engelmann.

Guopeng Yin, Qingyuan Zhang, and Yimeng Li. 2014. Effects of emotional valence and arousal on consumer perceptions of online review helpfulness. In *Twentieth Americas Conference on Information Systems*. Savannah, USA.