# Corrigendum to: Sentiment Lexica from Paired Comparisons

by Christoph Dalitz
Technical Report No. 2017-02, Hochschule Niederrhein,
Fachbereich Elektrotechnik und Informatik, 2017

and

# Sentiment Lexica from Paired Comparisons

by Christoph Dalitz & Katrin E. Bednarek
International Conference on Data Mining (ICDM)
Sentire Workshop, pp. 924-930, 2016

## Important notice

*A corrected and considerably extended version
of the ICDM paper has subsequently been published as*

C. Dalitz, J. Wilberg, K.E. Bednarek: „Paired Comparison Sentiment Scores.“ Technical Report No. 2017-03, Hochschule Niederrhein, Fachbereich Elektrotechnik und Informatik (2017)

# Corrigendum to: Sentiment Lexica from Paired Comparisons

Christoph Dalitz
Institut für Mustererkennung
Hochschule Niederrhein
Reinarzstr. 49, 47805 Krefeld
christoph.dalitz@hsnr.de

**Abstract**

The sentiment scores presented by Dalitz & Bednarek in "Sentiment lexica from paired comparisons" at the ICDM Sentire workshop (2016) were based on an approximation formula by Elo that was grossly inaccurate in that particular use case. This corrigendum describes how the scores should be estimated instead and shows that these new scores are indeed a good fit to the probabilistic sentiment score model. The conclusions in the Sentire paper about the quality of the corpus based sentiment lexica SentiWS and SenticNet 3 still hold, however, because the scores obtained with the inaccurate approximation formula are similar to the correctly estimated scores when scaled with a factor, which means that the correlation is not that much affected by the error. Nevertheless, the approximate solution presented in the Sentire paper should not be used and be replaced by a numerical non-linear least squares or maximum likelihood optimization.

## 1 Introduction

In their presentation at the ICDM Sentire workshop [1], Dalitz & Bednarek proposed a method to assign polarity scores to words that represents the strength of the positive or negative affect associated with each word. The method uses the paired comparisons, the theory of which was originally developed in psychology [2] and later applied to chess ratings [3, 4]. The original model ignored the possibility of draws, but Dalitz & Bednarek used a generalized model that allowed draws, too [5].

Applied to word polarity, the model makes the assumption that each word $w_i$ has a hidden rating $r_i$. The probability that $w_i$ is more positive than $w_j$ (symbolically: $w_i > w_j$) in a randomly chosen context depends on the difference between the hidden ratings:

$$P(w_i > w_j) = F(r_i - r_j - t) \tag{1a}$$

$$P(w_i \approx w_j) = F(r_i - r_j + t) - F(r_i - r_j - t) \tag{1b}$$

$$P(w_i < w_j) = F(r_j - r_i - t) \tag{1c}$$

where $(-t, t)$ is the *draw width*, and $F$ is the cumulative distribution function of a zero-symmetric random variable. The normal distribution function is the only choice for $F$ with a sound statistical justification (Thurstone-Mosteller model), but simpler forms for $F$ have also been used like the logistic distribution (Bradley-Terry model) or the uniform distribution [6].

The ratings are the sentiment scores and need to be estimated from the observed comparison results. To estimate all scores, we performed a $k$-fold round-robin experiment from which the $n$ unknown scores $(r_i)_{i=1}^n$ were to be estimated (case 2 in [1]). To do so, we followed the non-linear least squares estimation method by Batchelder & Bershed [3], which minimizes the squared differences between the observed scorings

$$S_i = \underbrace{W_i}_{\text{wins}} + \frac{1}{2}(\underbrace{D_i}_{\text{draws}} + \underbrace{k}_{\text{self}}) \tag{2}$$

and and their expectation values $E(S_i)$, which can be approximated by a Taylor expansion around $t = 0$ as

$$E(S_i) = k \sum_{j=1}^{n} F(r_i - r_j) + O(t^2) \tag{3}$$

The non-linear least squares estimator are the ratings $r_1, \ldots, r_n$ that minimize

$$SS(r_1, \ldots, r_n) = \sum_{i=1}^{n} \Big( S_i - k \sum_{j=1}^{n} F(r_i - r_j) \Big)^2 \tag{4}$$

In [1], we had solved Eq. (4) for its minimum analytically by making the following approximation that is due to Elo [4, paragraph 1.66]:

$$\sum_{j=1}^{n} F(r_i - r_j) \approx n \cdot F(r_i - \overline{r}) \tag{5}$$

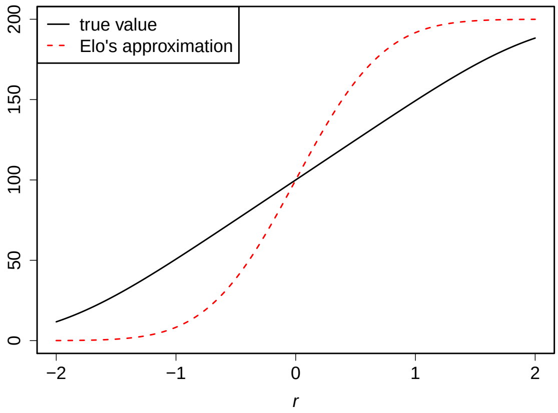


**Figure 1:** Comparison of Elo's approximation $n \cdot F(r - \overline{r})$ (see Eq. (5)) with the true value of $\sum_{i=1}^{n} F(r - r_i)$ as a function of $r$ for evenly spaced $r_i$ and a normal distribution $F$ with $\sigma = 1/\sqrt{3}$.

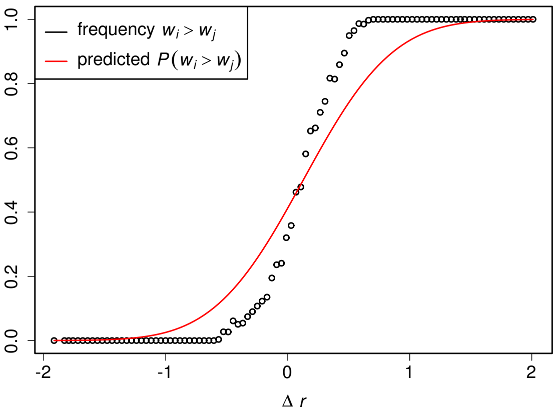


**Figure 2:** Comparison of the observed relative frequencies with the probabilities predicted by model (1) with a normal distribution $F$ and scores computed with Elo's approximation.

where $\overline{r} = \sum_{j=1}^{n} r_i/n$ is the average rating of all words. Eq. (5) holds exactly only when $F$ is the uniform distribution and all rating differences are within the support of the uniform distribution. In all other cases, the approximation may become inaccurate. As we will show in this corrigendum, the error is so large in this use case that the approximation must not be used and a numeric algorithm for minimizing $SS(r_1, \ldots, r_n)$ must be applied instead.

## 2 Inaccuracy of Elo's approximation

Let us first check directly how good Elo's approximation is in our case. We will see in the next section that the resulting scores range in our experiment is about $[-2, +2]$ for a normal distribution function $F$ with $\sigma = 1/\sqrt{3}$. For $n = 200$ ratings equally spaced between $-2$ and $+2$, the values of the different sides of Eq. (5) are shown in Fig. 1.

The difference can become greater than 40 which is an error of 20% of the total possible score 200. This shows that Elo's approximation is too crude to be usable in our case. It is interesting to note that the true value is close to a linear function of $r$ and one could get the idea to use this approximation. The slope of the line depends on the range of the scores, however, which is not known beforehand. This means that a linear approximation of $\sum_{i=1}^{n} F(r - r_i)$ cannot be used either to find the minimum of (4) analytically.

Another way to assess the quality of the approximation (5) is to compare the observed probabilities with the probabilities predicted by the model (1) with the scores obtained from the approximation through [1, Eq. (9)]. To do so, we have binned all occurring $\Delta r = r_i - r_j$ in our 2-fold round-robin experiment into 100 bins and counted the relative frequencies of $w_i > w_j$, $w_i \approx w_j$, and $w_i < w_j$ as estimators for the respective probabilities. Fig. 2 shows that the prediction is quite poor and that the scores presented at the Sentire workshop are not the best fit to the model because the score differences are estimated too small.

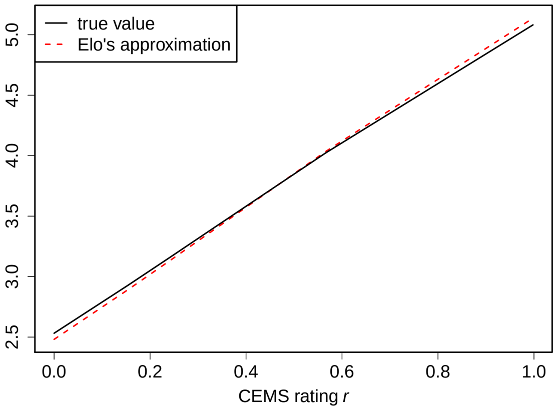


**Figure 3:** Comparison of Elo's approximation $n \cdot F(r - \overline{r})$ with the true value of $\sum_{i=1}^{n} F(r - r_i)$ for the CEMS data with the scores reported by Cattelan [5] for the normal distribution $F$ and $\sigma = 1$.

This raises the question why Elo's approximation worked so well in the case of the CEMS data [7], where it yielded almost the same results as the maximum-likelihood estimator. As shown in Fig. 3, the ratings are so close in this case that all differences

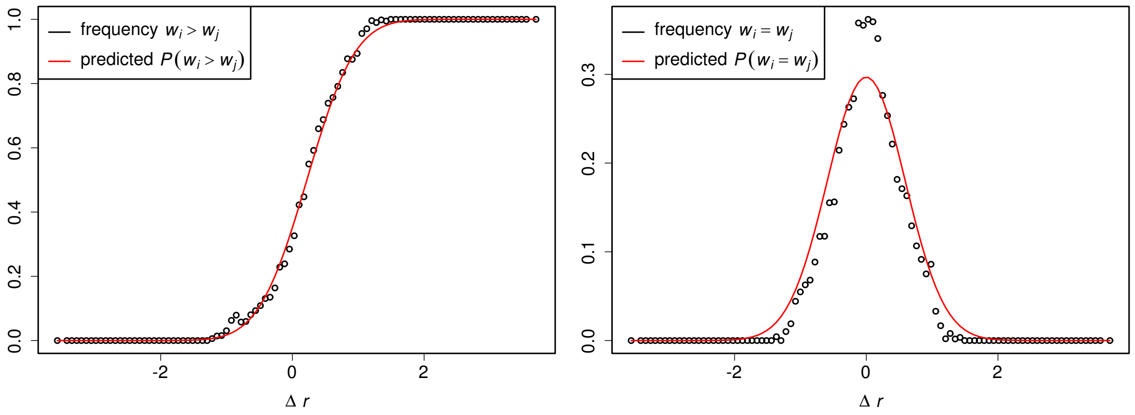


**Figure 4:** Comparison of the observed relative frequencies with the probabilities predicted by model (1) with a normal distribution $F$ and scores computed by numeric minimization of (4).

fall into a range where the distribution function is almost linear. This means that Elo's approximation incidentally is quite good in this particular case. For our word sentiment score problem, however, a different estimation method must be used.

## 3 Correct score estimation

For obtaining correct scores, the sum of squares (4) must be minimized numerically. For non-linear least squares problems, the Levenberg-Marquardt algorithm is an efficient algorithm that is, e.g., provided by the R package *minpack.lm* [8]. As can be seen in Fig. 4, the ratings estimated with this method lead to a model in good agreement with the observed comparison results.

The range of the least squares fitted scores is about $[-2, 2]$ when $\sigma$ is set to $1/\sqrt{3}$, while the range of the scores obtained with Elo's approximation was about $[-1, 1]$. The draw width $t$ is greater, too (0.220 versus 0.128). Nevertheless, the Pearson correlation between both scores is 0.9973, and the Spearman correlation even 1.0000. This means that the scores from Elo's approximation are linearly transformed by a factor around 0.5, which theoretically could be corrected *after* score estimation by reducing the scale parameter $\sigma$ in $F$. This means that the probability $P$("unpraktisch" $>$ "rüde") reported in [1, p. 928] was too small (0.58) and is actually greater (0.64).

An alternative approach to estimate $r_1, \ldots, r_n$ would be to maximize the log-likelihood function

$$l(r_1, \ldots, r_n, t) = \sum_{\substack{\textit{comparisons}\\ \textit{with } w_i > w_j}} \log F(r_i - r_j - t) \qquad (6)$$

$$+ \sum_{\substack{\textit{comparisons}\\ \textit{with } w_i \approx w_j}} \log \Big( F(r_i - r_j + t) - F(r_i - r_j - t) \Big)$$

The resulting model fit is similar to Fig. 4, but with an even slightly wider range of score values: $[-1.955, 2.132]$ versus $[-1.832, 1.904]$. The runtime for maximum-likelihood estimation is considerably greater[1], however, and numeric optimization of (6) fails in the case of the uniform distribution, because the objective function is not differentiable and many values of the ratings lead to zero probabilities. The best fit with non-linear least squares even has $l(r_1, \ldots, r_n, t) = -\infty$. For other than the uniform distribution, the maximum-likelihood estimation is a good alternative, however, especially as it does not make the assumption of a small draw width $t$. In our situation it is $t \approx 0.2$, and the Taylor expansion around $t = 0$ is justified, but this might not hold in more general use cases of the paired comparison model.

## 4 Correlation with other lexica

In the presentation for the Sentire workshop, we had used the scores to evaluate the relative quality of the corpus based sentiment lexica SentiWS [9] and SenticNet 3 [10] by means of their Pearson correlation with the paired comparison scores. Based on these correlations, we concluded that SenticNet is in better agreement with our ground truth data. As can be seen from Table 1, our conclusion still holds with the

[1] Numeric minimization of (4) with the R function *nls.lm* took 12s on an Intel i7-4770, while it took 8min for the maximization of (6) with *optim*.

| | *choice for $F$* | | | |
|---|---|---|---|---|
| | *normal* | *logistic* | *uniform* | |
| *direct* | 0.977 | 0.978 | 0.973 | *LSQ* |
| *SentiWS* | 0.714 | 0.715 | 0.713 | |
| *SenticNet* | 0.759 | 0.762 | 0.751 | |
| *direct* | 0.968 | 0.961 | 0.979 | *Elo* |
| *SentiWS* | 0.709 | 0.707 | 0.710 | |
| *SenticNet* | 0.741 | 0.732 | 0.763 | |

**Table 1:** Pearson correlation $r_p$ of the polarity scores with scores from direct assignment and corpus-based lexica. "LSQ" are the results with scores correctly estimated with non-linear least squares. For comparison, the correlations with the erroneously estimated scores from [1] are given ("Elo").

correctly estimated scores, although they have a range about twice as wide. As the correlation between the erroneous (Elo) and the correct (LSQ) scores is high, the difference in their range has less effect on their correlation with other sentiment lexica than one should have expected from the inaccuracy of Elo's approximation in this case.

There is one notable difference, however: for the correct scores, the uniform distribution no longer shows the highest correlation with the other lexica. On the contrary: it is lower, albeit only slightly. Moreover, the plot for the uniform distribution corresponding to Fig. 4 shows a slightly poorer agreement between model prediction and observed judgments. In contrary to the suggestion in [1], there is thus no reason to prefer the uniform distribution.

## 5 Conclusion

The approximation formula for estimating the word sentiment scores in [1] must not be used. The scores must instead be computed either by non-linear least squares minimization of Eq. (4), or by maximizing the log-likelihood function (6). This also affects the computation of scores for new words, where the estimation step in lines 20 and 24 of Algorithm 1 [1, p. 927] must be replaced with a maximum likelihood or non-linear least squares estimate.

A more general lesson can be learned from this example: always verify the approximations made in a model after fitting the model to the observed data! I am sorry that we did not do this before our Sentire presentation and that this corrigendum was necessary.

## References

[1] C. Dalitz and K. E. Bednarek, "Sentiment lexica from paired comparisons," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 924–930, 2016.

[2] L. L. Thurstone, "A law of comparative judgment.," *Psychological Review*, vol. 34, no. 4, pp. 368–389, 1927.

[3] W. H. Batchelder and N. J. Bershad, "The statistical analysis of a Thurstonian model for rating chess players," *Journal of Mathematical Psychology*, vol. 19, no. 1, pp. 39–60, 1979.

[4] A. E. Elo, *The Rating of Chess Players, Past and Present*. New York: Arco, 1978.

[5] M. Cattelan, "Models for paired comparison data: A review with emphasis on dependent data," *Statistical Science*, vol. 27, no. 3, pp. 412–433, 2012.

[6] G. E. Noether, "Remarks about a paired comparison model," *Psychometrika*, vol. 25, no. 4, pp. 357–367, 1960.

[7] R. Dittrich, R. Hatzinger, and W. Katzenbeisser, "Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 4, pp. 511–525, 1998.

[8] T. V. Elzhov, K. M. Mullen, A.-N. Spiess, and B. Bolker, *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK*, 2016. R package version 1.2-1.

[9] R. Remus, U. Quasthoff, and G. Heyer, "SentiWS - a publicly available German-language resource for sentiment analysis," in *Conference on Language Resources and Evaluation (LREC)*, pp. 1168–1171, 2010.

[10] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," in *AAAI conference on artificial intelligence*, pp. 1515–1521, 2014.

# Sentiment Lexica from Paired Comparisons

Christoph Dalitz and Katrin E. Bednarek
Institute for Pattern Recognition
Niederrhein University of Applied Sciences,
Reinarzstr. 49, 47805 Krefeld, Germany
Email: christoph.dalitz@hsnr.de

***Abstract*—The method of paired comparison is an established method in psychology for assigning ranks or inherent score values to different stimuli. This article describes how this method can be used for building a sentiment lexicon and for extending the lexicon with arbitrary new words. An initial lexicon with $n = 200$ German words is created from a two-fold all-pair comparison experiment with ten different test persons. A cross-validation experiment suggests that only two-fold $\log_2(n)+8 = 16$ comparisons are necessary to estimate the score of a new, yet unknown word. We make the new lexicon available and compare it with the corpus-based lexica SentiWS and SenticNet.**

## I. Introduction

A *sentiment lexicon* is a dictionary that assigns each term a *polarity score* representing the strength of the positive or negative affect associated with the term. In general, word polarity strength depends on the context, and its representation by a single number can therefore only be a crude approximation. Nevertheless, such sentiment lexica are an important tool for opinion mining and have been proven to be very useful. Examples for recent use cases are the sentiment analysis of tweets and SMS [1] or the political classification of newspapers [2].

There are two approaches to building a sentiment lexicon: *corpus based* automatic assignment or *manual annotation*. Corpus based approaches start with a set of seed words of known polarity and extend this set with other words occurring in a text corpus or a synonym lexicon. One possible approach is to compute the "Pointwise Mutual Information" (PMI) [3] from cooccurrences of seed words and other words. The German sentiment lexicon *SentiWS* [4] was built in this way. A more sophisticated corpus-based method was implemented for *SenticNet* [5], [6]. Such methods can even be extended to automatically assign emotion categories to terms [7].

Corpus based methods have the advantage of building large lexica in an automated way without time consuming experiments with human annotators. They have two drawbacks, however: due to peculiarities in the corpus, some words can obtain strange scores. In SentiWS 1.8, e.g., "gelungen" (*successful*) has the highest positive score (1.0) while the more positive word "fantastisch" (*fantastic*) only has a score of 0.332. In SenticNet 3.0, "inconsequent" has a strong positive polarity (0.948). Moreover, it is not possible to assign a score value to words that are absent from the corpus.

Assigning polarity scores by manual annotations can be done in two different ways. One is by direct assignment of an ordinal score to each word on a coarse scale. In this way, Wilson et al. have created a subjectivity lexicon with English words [8], which has also been used by means of automated translations for sentiment analysis of German texts [9]. The other method is to present words in pairs and let the observer decide which word is more positive or more negative. Comparative studies for other use cases have shown that scores from paired comparisons are more accurate than direct assignments of scores [10]. The main advantage is their invariance to scale variances between different test persons. This is especially important when words are added at some later point when the original test persons are no longer available. Unfortunately, paired comparisons are much more expensive than direct assignments: for $n$ words, direct assignments only require $O(n)$ judgments, while a complete comparison of all pairs requires $O(n^2)$ judgments. For large $n$, this becomes prohibitive and must be replaced by incomplete comparisons, i.e. by omitting pairs. Incomplete paired comparisons are widely deployed in the estimation of chess players' strength [11], [12].

In the present paper, we propose a method for building a sentiment lexicon from paired comparisons in two steps. At first, an initial lexicon is built from a limited set of 200 words by comparison of all pairs. This lexicon is then subsequently extended with new words, which are only compared to a limited number of words from the initial set, which are determined based on Silverstein & Farrell's sorting method [13]. Sec. II provides an overview over the mathematical methods of the method of paired comparisons, Sec. III describes the criteria for choosing the initial set of words and our experimental setup, and Sec. IV presents the results for the initial lexicon, compares it to SentiWS and SenticNet, and evaluates a method for adding new words. The new lexicon will be made available on the authors' website.

## II. Method of paired comparison

The method of paired comparison goes back to the early 20th century [14]. See [12] for a comprehensive presentation of the model and its estimation problems, and [15] for a review of recent extensions. Applied to word polarity, it makes the assumption that each word $w_i$ has a hidden score (or rating) $r_i$. The probability that $w_i$ is more positive than $w_j$ (symbolically: $w_i > w_j$) in a randomly chosen context depends on the difference between the hidden scores:

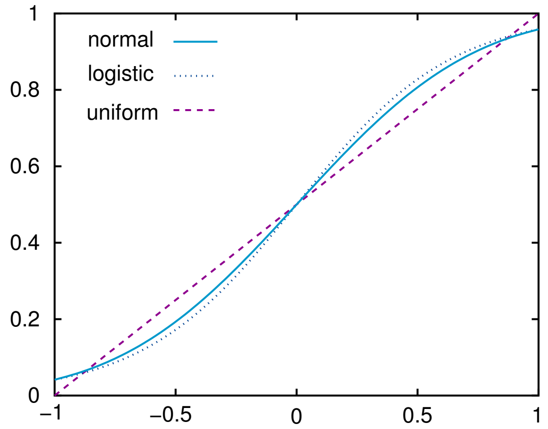


Fig. 1. Different choices for the cumulative distribution function $F$ with identical standard deviations $\sigma = 1/\sqrt{3}$.

$$P(w_i > w_j) = F(r_i - r_j - t) \tag{1a}$$

$$P(w_i \approx w_j) = F(r_i - r_j + t) - F(r_i - r_j - t) \tag{1b}$$

$$P(w_i < w_j) = F(r_j - r_i - t) \tag{1c}$$

where $(-t, t)$ is the *draw width*, and $F$ is the cumulative distribution function of a zero-symmetric random variable. Thurstone's model [14] uses an $F$ based on the normal distribution, a model that can be derived from the assumption that the polarity of a word $w_i$ is normally distributed around its mean inherent score $r_i$. Although this is the only model with a sound statistical justification, simpler distribution functions have also been used for convenience, e.g. the logistic distribution (Bradley-Terry model) or the uniform distribution, which is the only one which strictly limits the range of the rating differences $r_i - r_j$ (see Fig. 1). The standard deviation $\sigma$ of the distribution function is a scale parameter that determines the range of the ratings $r_i$.

As the probabilities in Eq. (1) only depend on rating differences, the origin $r = 0$ cannot be determined from the model, but must be defined by an external constraint. Typical choices are the average rating constraint $\sum_i r_i = 0$, or the reference object constraint, i.e. $r_i = 0$ for some $i$. For sentiment lexica, a natural constraint can be obtained by separately classifying words into positive and negative words and choosing the origin in such a way that the scores from the paired comparison model coincide with these classifications.

The ratings $r_i$ and the draw-width $t$ must be estimated from the observed comparisons. During our two steps of building a sentiment lexicon, two different estimation problems occur:

1) Estimation of one unknown $r$ of a new word from $m$ comparisons with old words with known ratings $q_i, \ldots, q_m$.
2) Estimation of $t$ and all unknown $r_1 \ldots, r_n$ from round-robin pair comparisons.

Estimators with desirable properties are generally obtained from maximizing the (log) likelihood function, which can only be done numerically in the above cases. Alternatively, approximate analytic formulas for estimating the parameters can be obtained with the "generalized method of moments" as outlined in the following two subsections.

### A. Case 1: one unknown rating $r$

Let us first consider this simpler case. The idea of the generalized method of moments is to set the measured value of an observable equal to its expectation value and solve the resulting equation for the parameters. Following [12], we choose as an observable a combination of the number of wins $W$ of the new word and the number of draws $D$, which we set equal to its expectation values

$$W = \sum_{i=1}^{m} F(r - q_i - t) \tag{2a}$$

$$D = \sum_{i=1}^{m} \Big( F(r - q_i + t) - F(r - q_i - t) \Big) \tag{2b}$$

For small $t$, we can make a Taylor expansion of the right hand sides of Eq. (2) around $t = 0$, and, for the combination $W + D/2$, the term linear in $t$ vanishes:

$$W + D/2 \approx \sum_{i=1}^{m} F(r - q_i) \tag{3}$$

With Elo's approximation[1] $\sum_{i=1}^{m} F(r - q_i) \approx m \cdot F(r - \overline{q})$ [11], this can be solved for $r$ in closed form:

$$r \approx \overline{q} + F^{-1}\Big(\frac{W + D/2}{m}\Big) \quad \text{with} \quad \overline{q} = \frac{1}{m}\sum_{i=1}^{m} q_i \tag{4}$$

An alternative solution can be obtained by numerically maximizing the log-likelihood function $l(r)$ ($t$ is considered as given):

$$\begin{aligned} l(r) = & \sum_{\mathit{wins}} \log F(r - q_i - t) \\ & + \sum_{\mathit{draws}} \log \Big( F(r - q_i + t) - F(r - q_i - t) \Big) \\ & + \sum_{\mathit{losses}} \log F(q_i - r - t) \end{aligned} \tag{5}$$

### B. Case 2: all ratings $(r_i)_{i=1}^n$ and $t$ unknown

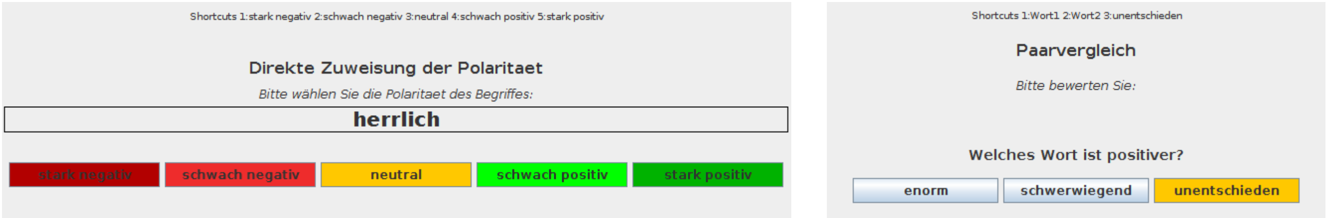
Again, we obtain an approximate estimator with the generalized method of moments by considering for each word $w_i$ the total score $S_i$ from $k$-fold round-robin comparisons as an observable

$$S_i = \underbrace{W_i}_{\text{wins}} + \frac{1}{2}(\underbrace{D_i}_{\text{draws}} + \underbrace{k}_{\text{self}}) \tag{6}$$

and setting it equal to its expectation value. With a Taylor approximation around $t = 0$ and Elo's approximation, we obtain

$$S_i \approx k \sum_{j=1}^{n} F(r_i - r_j) \approx kn\, F(r_i - \overline{r}) \tag{7}$$

[1] This holds exactly for the uniform distribution, but is only a crude approximation for the Thurstone or Bradley-Terry model.


Shortcuts 1:stark negativ 2:schwach negativ 3:neutral 4:schwach positiv 5:stark positiv

Direkte Zuweisung der Polaritaet

*Bitte wählen Sie die Polaritaet des Begriffes:*

**herrlich**

Shortcuts 1:Wort1 2:Wort2 3:unentschieden

Paarvergleich

*Bitte bewerten Sie:*

Welches Wort ist positiver?

(a) direct assignment

(b) paired comparison

Fig. 2. Graphical user interface for score assignment as seen by the test persons.

where $\overline{r} = \sum_j r_j/n$ is the average rating of all words. Joint estimates for all ratings can then be obtained by minimizing the sum of the squared deviations

$$SS(r_1,\dots,r_n) = \sum_{i=1}^{n} \Big(S_i - kn\,F(r_i - \overline{r})\Big)^2 \qquad (8)$$

The minimum of expression (8) can be given in closed form because (8) is exactly zero for (note that $\overline{r}$ can be chosen arbitrarily, as explained in section II):

$$r_i = \overline{r} + F^{-1}(S_i/kn) \quad \text{with} \quad \overline{r} = \frac{1}{n}\sum_{j=1}^{n} r_j \qquad (9)$$

To obtain an approximate estimator for the draw width $t$, let us consider the total number of draws $D_i$ of each word $w_i$ as an observable and set it equal to its expectation value in a $k$-fold round robin experiment:

$$D_i = k\sum_{j\neq i}\Big(F(r_i - r_j + t) - F(r_i - r_j - t)\Big) \qquad (10)$$

Keeping only the first non-zero term in a Taylor expansion around $t = 0$ of the sum on the right hand side yields

$$\sum_{j\neq i}\Big(F(r_i{-}r_j{+}t){-}F(r_i{-}r_j{-}t)\Big) \approx 2t\sum_{j\neq i}F'(r_i{-}r_j) \quad (11)$$

Again, we can determine $t$ by minimizing the sum of the squared deviations

$$SS(t) = \sum_{i=1}^{n}\left(D_i - 2kt\sum_{j\neq i}^{n}F'(r_i - r_j)\right)^2 \qquad (12)$$

The minimum of expression (12) can be found analytically by solving for the zero of $SS'(t)$, which yields

$$t = \frac{\sum_{i=1}^{n} f_i D_i/2}{\sum_{i=1}^{n} f_i^2} \quad \text{with} \quad f_i = k\sum_{j\neq i}F'(r_i - r_j) \qquad (13)$$

The approximate solution (9) and (13) can then be used as a starting point for maximizing the log-likelihood function

$$l(r_1,\dots,r_n,t) = \sum_{\substack{\textit{comparisons}\\ \textit{with } w_i > w_j}} \log F(r_i - r_j - t) \qquad (14)$$
$$+ \sum_{\substack{\textit{comparisons}\\ \textit{with } w_i \approx w_j}} \log\Big(F(r_i - r_j + t) - F(r_i - r_j - t)\Big)$$

It should be noted that, due to the large number of $n+1$ parameters, numerical methods for maximizing (14) might not work reliably. In this case, the approximate solution (9) and (13) should be used.

## III. Experimental design

To select 200 words for building the initial lexicon from round robin pair comparisons, we have started with all 1 498 adjectives from SentiWS [4]. To build an intersection of these words with SenticNet [5], we translated all words into English with both of the German-English dictionaries from *www.dict.cc* and *www.freedict.org*, and removed all words without a match in SenticNet. From the remaining 1 303 words, we selected manually 10 words that appeared strongly positive to us, and 10 strongly negative words. This was to make sure that the polarity range is sufficiently wide in the initial lexicon. The remaining words were ranked by their SentiWS score and selected with equidistant ranks, such that we obtained 200 words, with an equal number of positive and negative words according to SentiWS.

We then let ten different test persons assign polarity scores to these words in two different experiments. The first one consisted of direct assignment of scores on a five degree scale (see Fig. 2(a)), which resulted in ten evaluations for each word. An average score was computed for each word by replacing the ordinal scale with a metric value ($-1$ = strong negative, $-0.5$ = weak negative, $0$ = neutral, $0.5$ = weak positive, 1.0 = strong positive).

The second experiment consisted of twofold round robin paired comparisons, with all $2\cdot 19\,900$ pairs evenly distributed among the ten test persons, such that each person evaluated 3 980 pairs. See Fig. 2(b) for the graphical user interface presented to the test persons. The scores were computed with the method-of-moments solution from section II-B. The standard deviation of the normal distribution was set to $\sigma = 1/\sqrt{3}$, which corresponds to the distribution function in Fig. 1. For a reasonable choice for the origin $r = 0$, we shifted all scores such that they best fitted to the discrimination between positive and negative words from the direct comparison experiment. To be precise: when $r_i'$ is the score from the direct assignment and $r_i$ the score from the paired comparisons with an arbitrarily set origin, we chose the shift value $\rho$ that minimized the squared error

$$SE(\rho) \quad = \sum_{\mathrm{sign}(\rho+r_i)\neq\mathrm{sign}(r_i')} (\rho + r_i)^2 \qquad (15)$$

**Algorithm 1** One-fold addition of new word

```
Input: word $w$ with unknown rating $r$, words
    $\vec{v} = (v_1, \ldots, v_n)$ sorted by their known ratings $q_1, \ldots, q_n$
Output: new rating $r$
$i_l \leftarrow 1$ and $i_r \leftarrow n$
$i \leftarrow \lfloor (i_l + i_r)/2 \rfloor$
$m_0 \leftarrow 0$
$\vec{q} \leftarrow ()$
$\vec{u} \leftarrow \vec{v}$
while $i > i_l$ and $i < i_r$ do                              ▷ binary search
    $m_0 \leftarrow m_0 + 1$
    $\vec{q} \leftarrow \vec{q} \cup q_i$
    $s \leftarrow$ score from $w$ versus $v_i$ comparison,
        where win counts 1 and draw counts 1/2
    $S \leftarrow S + s$
    $\vec{u} \leftarrow \vec{u} \setminus v_i$
    if $s > 1/2$ then
        $i_l \leftarrow i$
    else
        $i_r \leftarrow i$
    end if
    $i \leftarrow \lfloor (i_l + i_r)/2 \rfloor$
end while
$r_0 \leftarrow$ mean$(\vec{q}) + F^{-1}(S/m_0)$                ▷ first guess
$\vec{u} \leftarrow m$ words in $\vec{u}$ with closest ratings to $r_0$
$\vec{q} \leftarrow \vec{q} \cup$ ratings of $\vec{v}$
$S \leftarrow S+$ total score of $w$ against words from $\vec{u}$
$r \leftarrow$ mean$(\vec{q}) + F^{-1}(S/(m_0 + m))$           ▷ cf. Eq. (4)
return $r$
```

For adding new words, we implemented the method by Silverstein & Farrell, which uses comparison results to sort the new word into a binary sort tree built from the initial words [13]. For $n$ initial words, this only leads to $\log_2(n)$ comparisons, which generally are too few for computing a reliable score. We therefore extended this method by adding comparisons with the $m$ words from the initial set which have the closest rank to the rank obtained from the sort tree process. Algorithm 1 lists the resulting algorithm in detail. This algorithm can be applied sequentially to more than one test person by estimating the resulting rating from all scores obtained from all test persons with Eq. (4). We have evaluated this method with a leave-one-out experiment using the comparisons from our two-fold round-robin comparison experiment.

## IV. Results

### A. Score values

It turned out that all maximization algorithms provided by the *R*-package *optimx* failed to maximize the log-likelihood function (14). We therefore used the approximate solution given by (9) and (13). To get an idea of the difference between both solutions, we compared them for a well-studied much smaller paired-comparison experiment, the student preference data for the Community of European Management Schools (CEMS) [16]. The data is available in the *R* Package *BradleyTerry2*[2] and was also used as an example in the review by Cattelan [15]. Theoretically, it should include all-pair preference choices between six management schools made by 303 students, but as 91 students missed answering some questions, it actually only includes 212 students performing a full round-robin comparison. This means that we effectively only have a 212-fold round-robin experiment.

| | *212 round-robin* | | | | *all 303* |
|---|---|---|---|---|---|
| | *MM/Elo* | | *ML* | | *ML* |
| *school* | $r_i$ | $\sigma_{JK}$ | $r_i$ | $\sigma_{JK}$ | $r_i$ |
| London | 0.555 | 0.038 | 0.632 | 0.046 | 0.588 |
| Paris | 0.177 | 0.045 | 0.193 | 0.050 | 0.156 |
| Barcelona | -0.047 | 0.042 | -0.064 | 0.046 | -0.078 |
| St.Gallen | -0.120 | 0.046 | -0.121 | 0.051 | -0.086 |
| Milano | -0.147 | 0.041 | -0.176 | 0.045 | -0.169 |
| Stockholm | -0.417 | 0.039 | -0.465 | 0.044 | -0.410 |
| $t$ | 0.162 | 0.016 | 0.166 | 0.016 | 0.153 |

TABLE I
CEMS preference scores from method-of-moments (MM/ELo) and maximum-likelihood (ML).

We have computed the rating estimators from these 212 students both with the approximate method-of-moments and maximum-likelihood, and estimated the standard error with the jackknife variance $\sigma^2_{JK}$ [17] by cyclic omission of one student. All ratings were normalized to zero mean, and $F$ was chosen as a standard normal distribution[3]. The results are listed in Table I together with the maximum-likelihood estimators obtained from all 303 students including those students with missing answers in the last column. The difference between the different estimators is smaller than the estimated standard error in most cases, with the method-of-moments estimator surprisingly even closer on average to the estimator in the last column. We therefore conclude that the approximate method-of-moments estimators works well for estimating ratings from round-robin comparisons.

For the 200 words, we estimated the polarity ratings with the approximate method-of-moments with the three distribution functions of Fig. 1. The draw width $t$ turned out to be 0.128 for the normal distribution, 0.119 for the logistic distribution, and 0.146 for the uniform distribution. Fig. 3 shows a kernel density plot [18] for the resulting score distributions. The valley around zero (neutrality) is due to the fact that the words were drawn from the SentiWS data which only contains positive or negative words. The comparative shapes are as expected from Fig. 1: the steeper the slope of the distribution $F(x)$ at $x = 0$, the more condensed are the resulting scores.

It is interesting to compare the scores from paired comparisons for words which have obtained the same score from direct assignment on the five grade scale. The examples in table II show that the paired comparisons indeed lead to a different and finer rating scheme than averaging over coarse polarity

[2] http://cran.r-project.org/package=BradleyTerry2

[3] The choice $\sigma = 1$ was made for compatibility with the results reported by Cattelan in [15], which are identical to the last column in Table I when normalized to zero mean instead of zero minimum.

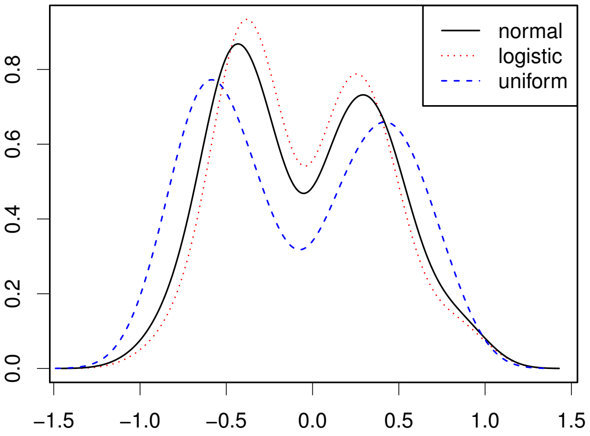


Fig. 3. Kernel density plot of the polarity score distribution in our sentiment lexicon for different cumulative distribution functions $F$.

| *adjective* | $r_{direct}$ | $r_{paired}$ | $\sigma_{JK}$ |
|---|---|---|---|
| paradiesisch (*paradisaical*) | 1.00 | 0.872 | 0.074 |
| wunderbar (*wonderful)* | 1.00 | 0.816 | 0.071 |
| perfekt (*perfect*) | 0.95 | 1.001 | 0.086 |
| traumhaft (*dreamlike*) | 0.95 | 0.955 | 0.081 |
| prima (*great)* | 0.75 | 0.684 | 0.063 |
| zufrieden (*contented*) | 0.75 | 0.495 | 0.055 |
| kinderleicht (*childishly simple*) | 0.50 | 0.348 | 0.051 |
| lebensfähig (*viable*) | 0.50 | 0.249 | 0.048 |
| ausgeweitet (*expanded*) | 0.05 | -0.008 | 0.046 |
| verbindlich (*binding*) | 0.00 | 0.091 | 0.039 |
| kontrovers (*controversial*) | -0.05 | -0.175 | 0.047 |
| unpraktisch (*unpractical*) | -0.50 | -0.279 | 0.046 |
| rüde (*uncouth*) | -0.50 | -0.517 | 0.052 |
| falsch (*wrong*) | -0.75 | -0.515 | 0.055 |
| unbarmherzig (*merciless*) | -0.75 | -0.688 | 0.055 |
| erbärmlich (*wretched*) | -1.00 | -0.728 | 0.055 |
| tödlich (*deadly*) | -1.00 | -1.028 | 0.062 |

TABLE II
EXAMPLE SCORES FROM AVERAGE DIRECT ASSIGNMENT AND PAIRED COMPARISONS WITH THE NORMAL DISTRIBUTION.

scores from direct assignments, and that they also can lead to a reversed rank order (see, e.g., “traumhaft” and “wunderbar”). We have also estimated the variances of the polarity score estimates as the jackknife variance $\sigma^2_{JK}$ via cyclic omission of one word. These can be used to test whether, for $r_i > r_j$, the score difference is significant by computing the $p$-value $1 - \Phi\left((r_i - r_j)/\sqrt{\sigma_i^2 + \sigma_j^2}\right)$, where $\Phi$ is the distribution function of the standard normal distribution. For the words “unpraktisch” and “rüde”, e.g., the $p$-value is 0.0003, which is smaller than 5% and the difference is therefore statistically significant. The probability that “unpraktisch” is considered less negative than “rüde” is $F(-0.279-(-0.517)-0.128) = 0.58$.

### B. Adding new words

To obtain a lower bound for the error in estimating scores for unknown words, we have first computed the scores for all words with the estimators for one unknown rating $r$ as described in section II-A, where each word was compared with all other words and the scores $q_i$ for other words were considered to be known from the results in the preceding section. The mean absolute error with respect to the known score was much higher for the maximum-likelihood estimator (0.1444) than for method-of-moments estimator (0.008). This does not necessarily mean that the method-of-moments estimator is better, but it may be due to the fact that the “ground truth score” was also computed with the method-of-moments based on a similar observable. We therefore have used the method-of-moments estimator in the subsequent evaluations.

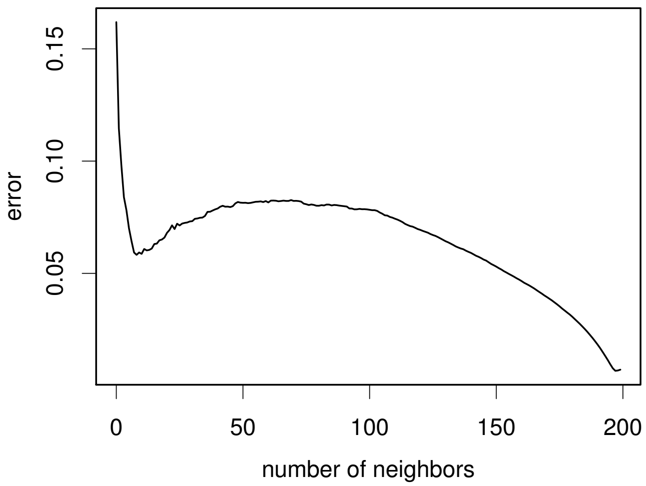


Fig. 4. Mean absolute error (MAE) from leave-one-out as a function of the number of additional comparisons after Silverstein & Farrell’s method.

For a reasonable recommendation for the number of incomplete comparisons, we have varied the number $m$ of neighboring scores after sorting in the unknown word with Silverstein & Farrell’s method (see section III). The results are shown in Fig. 4. It is interesting to observe that adding comparisons with similar scores first improves the accuracy, but leads to slight deterioration when too many similar words are added. The local minimum in Fig. 4 occurs at $m = 8$ with a mean absolute error of 0.0582. This effect deserves further investigation. A possible explanation for this behavior could be that we only had two results for each comparison, which are not sufficiently representative for comparisons of words with similar scores. Nevertheless, adding similar words after a first guess based on Silverstein & Farrel’s method leads to a smaller error than choosing comparison words at random: in a 100-fold Monte-Carlo experiment with choosing $\log_2(n)+m \approx 16$ words at random, we obtained a mean absolute error of 0.0840.

It should be noted that the error of 0.0582 is close to the standard deviations for the scores given in Table II and is about half the draw width. We therefore conclude that incomplete comparisons with only 16 out of 200 words provides a reasonably accurate score estimate, provided the words are selected with our method.

| | *choice for F* | | |
|---|---|---|---|
| | *normal* | *logistic* | *uniform* |
| *direct* | $r_p = 0.968$ | $r_p = 0.961$ | $r_p = 0.979$ |
| *SentiWS* | $r_p = 0.709$ | $r_p = 0.707$ | $r_p = 0.710$ |
| *SenticNet* | $r_p = 0.741$ | $r_p = 0.732$ | $r_p = 0.763$ |

TABLE III
PEARSON CORRELATION $r_p$ OF THE PARITY SCORES FROM THE PAIRED COMPARISON WITH THAT OF DIRECT ASSIGNMENT AND CORPUS-BASED METHODS.

### C. *Comparison to corpus-based lexica*

The polarity scores computed in our experiments provide nice ground truth data for the evaluation of corpus-based polarity scores. We therefore compared the scores from SentiWS 1.8 and SenticNet 3.0 with the scores computed from test person answers. SenticNet only contains English words, from which we have computed scores for the German words by translating each German word with both of the German-English dictionaries from *www.dict.cc* and *www.freedict.org* and by averaging the corresponding scores.

A natural measure for the closeness between lists of polarity scores is Pearson's correlation coefficient $r_p$, which has the advantage that it is invariant both under scale and translation of the variables. This is crucial in our case, because score values from paired comparisons allow for arbitrary shift and scale as explained in section II. $r_p$ is highest for a linear relationship and smaller for other monotonous relationships. As can be seen in Table III, this means that its value depends on the shape of the model distribution function $F$. Whatever function is used, the correlation between the scores from direct assignment and paired comparison is very strong. This was to be expected, because both values stem from test persons.

The correlation with the paired scores is higher for SenticNet than for SentiWS. According to the significance tests in the R package *cocor* [19], this difference is not significant, however, on a 5% significance level. From the density plot in Fig. 5 and the scatter plots in Fig. 6, it is nevertheless easily understandable that SenticNet is slightly stronger correlated to the true polarity scores than SentiWS. As can be seen in Fig. 6, SentiWS has many identical scores with values $0.0040$ and $-0.0048$. This peculiar distribution of the SentiWS scores was also observed in the original paper presenting the SentiWS data set by Remus et al. (see Fig. 1 in [4]). The identical scores show up in Fig. 5 as a peak around neutrality, which corresponds to a *valley* (sic!) in the score distribution from paired comparisons. They do not have such a strong effect on the correlation coefficient $r_p$, because the identical values also lead to a lower standard deviation (0.32 for SentiWS versus 0.44 for SenticNet), which is part of the denominator of $r_p$. Based on these observations, we consider the polarity scores from SenticNet (via automatic translation) more reliable than the scores from SentiWS.

## V. CONCLUSIONS

The new sentiment lexicon from paired comparison is a useful resource that can be used for different aims. It can be used, e.g., as ground truth data for testing and comparing automatic corpus-based methods for building sentiment lexica, as we did in section IV-C. Or it can be used as a starting point for building specialized lexica for polarity studies. The method for adding new words makes the method of paired comparison applicable to studies with an arbitrary vocabulary because it yields accurate polarity scores even for rare words.

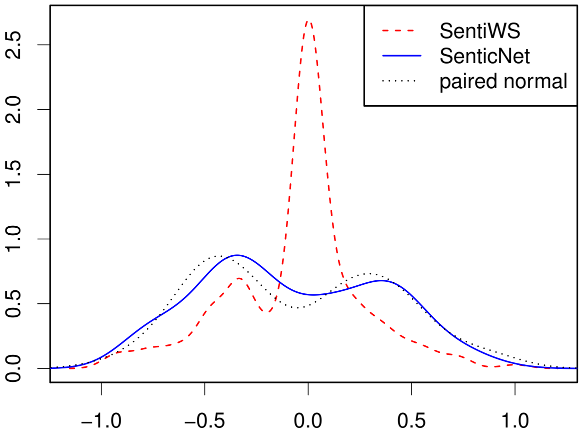


Fig. 5. Kernel density plots of the polarity score distributions.

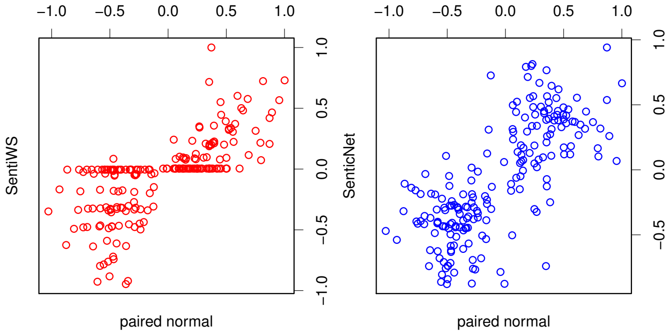


Fig. 6. Scatter plots comparing corpus-based scores with scores from paired comparisons.

Although the new sentiment lexicon is ready to be used, there are still two points in the method of paired comparison that require further research. One is the development of a robust numerical maximum-likelihood estimator that also works in the presence of draws and in the case of a large number of parameters. The other one is an explanation of the local minimum in Fig. 4: is this a general effect of our method for choosing words for comparison, or is it a peculiarity in our data?

The ratings presented in Table II have been calculated with the Thurstone model, which is the only model with a sound statistical justification. It might nevertheless be attractive in practice to use the uniform distribution, because it has a stronger correlation both with the scores from direct assignment and with the scores from SentiWS and SenticNet. Moreover it restricts the polarity scores to a limited range even in the presence of strongly positive or strongly negative words.

## ACKNOWLEDGMENTS

We are grateful to Jens Wilberg for implementing the translation based score computation for German words with SenticNet. Moreover, we thank the anonymous reviewers for their valuable comments.

## REFERENCES

[1] S. Kiritchenko, X. Zhu, and S. M. Mohammad, “Sentiment analysis of short informal texts,” *Journal of Artificial Intelligence Research*, pp. 723–762, 2014.

[2] K. Morik, A. Jung, J. Weckwerth, S. Rötner, S. Hess, S. Buschjäger, and L. Pfahler, “Untersuchungen zur Analyse von deutschsprachigen Textdaten,” Technische Universität Dortmund, Tech. Rep. 02/2015, 2015.

[3] K. W. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[4] R. Remus, U. Quasthoff, and G. Heyer, “SentiWS - a publicly available German-language resource for sentiment analysis,” in *Conference on Language Resources and Evaluation (LREC)*, 2010, pp. 1168–1171.

[5] E. Cambria, D. Olsher, and D. Rajagopal, “SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis,” in *AAAI conference on artificial intelligence*, 2014, pp. 1515–1521.

[6] E. Cambria and A. Hussain, *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Cham, Switzerland: Springer, 2015.

[7] S. Poria, A. Gelbukh, D. Das, and S. Bandyopadhyay, “Fuzzy clustering for semi-supervised learning–case study: Construction of an emotion lexicon,” in *Mexican International Conference on Artificial Intelligence*, 2012, pp. 73–86.

[8] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Conference on human language technology and empirical methods in natural language processing (HTL/EMNLP)*, 2005, pp. 347–354.

[9] M. Wiegand, C. Bocionek, A. Conrad, J. Dembowski, J. Giesen, G. Linn, and L. Schmeling, “Saarland University’s participation in the German sentiment analysis shared task (GESTALT),” in *Workshop Proceedings of the 12th KONVENS*, 2014, pp. 174–184.

[10] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, “Comparison of four subjective methods for image quality assessment,” *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[11] A. E. Elo, *The Rating of Chess Players, Past and Present*. New York: Arco, 1978.

[12] W. H. Batchelder and N. J. Bershad, “The statistical analysis of a Thurstonian model for rating chess players,” *Journal of Mathematical Psychology*, vol. 19, no. 1, pp. 39–60, 1979.

[13] D. A. Silverstein and J. E. Farrell, “Efficient method for paired comparison,” *Journal of Electronic Imaging*, vol. 10, no. 2, pp. 394–398, 2001.

[14] L. L. Thurstone, “A law of comparative judgment.” *Psychological Review*, vol. 34, no. 4, pp. 368–389, 1927.

[15] M. Cattelan, “Models for paired comparison data: A review with emphasis on dependent data,” *Statistical Science*, vol. 27, no. 3, pp. 412–433, 2012.

[16] R. Dittrich, R. Hatzinger, and W. Katzenbeisser, “Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 4, pp. 511–525, 1998.

[17] B. Efron and C. Stein, “The jackknife estimate of variance,” *The Annals of Statistics*, vol. 9, no. 3, pp. 586–596, 1981.

[18] S. Sheather and M. Jones, “A reliable data-based bandwidth selection method for kernel density estimation,” *Journal of the Royal Society series B*, vol. 53, pp. 683–690, 1991.

[19] B. Diedenhofen and J. Musch, “cocor: A comprehensive solution for the statistical comparison of correlations,” *PLoS ONE*, vol. 10, no. 4, p. e0121945, 2015.