Readers *vs.* Writers *vs.* Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation

Sven Buechel and **Udo Hahn**
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
{sven.buechel,udo.hahn}@uni-jena.de
http://www.julielab.de

Abstract

We here examine how different perspectives of understanding written discourse, like the reader's, the writer's or the text's point of view, affect the quality of emotion annotations. We conducted a series of annotation experiments on two corpora, a popular movie review corpus and a genre- and domain-balanced corpus of standard English. We found statistical evidence that the writer's perspective yields superior annotation quality overall. However, the quality one perspective yields compared to the other(s) seems to depend on the domain the utterance originates from. Our data further suggest that the popular movie review data set suffers from an atypical bimodal distribution which may decrease model performance when used as a training resource.

1 Introduction

In the past years, the analysis of subjective language has become one of the most popular areas in computational linguistics. In the early days, a simple classification according to the semantic polarity (positiveness, negativeness or neutralness) of a document was predominant, whereas in the meantime, research activities have shifted towards a more sophisticated modeling of sentiments. This includes the extension from only few basic to more varied emotional classes sometimes even assigning real-valued scores (Strapparava and Mihalcea, 2007), the aggregation of multiple aspects of an opinion item into a composite opinion statement for the whole item (Schouten and Frasincar, 2016), and sentiment compositionality on sentence level (Socher et al., 2013).

There is also an increasing awareness of different perspectives one may take to interpret written discourse in the process of text comprehension. A typical distinction which mirrors different points of view is the one between the writer and the reader(s) of a document as exemplified by utterance (1) below (taken from Katz et al. (2007)):

(1) Italy defeats France in World Cup Final

The emotion of the writer, presumably a professional journalist, can be expected to be more or less neutral, but French or Italian readers may show rather strong (and most likely opposing) emotional reactions when reading this news headline. Consequently, such finer-grained emotional distinctions must also be considered when formulating instructions for an annotation task.

NLP researchers are aware of this multi-perspectival understanding of emotion as contributions often target either one or the other form of emotion expression or mention it as a subject of future work (Mukherjee and Joshi, 2014; Lin and Chen, 2008; Calvo and Mac Kim, 2013). However, contributions aiming at quantifying the effect of altering perspectives are rare (see Section 2). This is especially true for work examining differences in annotation results relative to these perspectives. Although this is obviously a crucial design decision for gold standards for emotion analytics, we know of only one such contribution (Mohammad and Turney, 2013).

In this paper, we systematically examine differences in the quality of emotion annotation regarding different understanding perspectives. Apart from inter-annotator agreement (IAA), we will also look at other quality criteria such as how well the resulting annotations cover the space of possible ratings and check for the representativeness of the rating distribution. We performed a series of annotation experiments with varying instruc-

Proceedings of the 11th Linguistic Annotation Workshop, pages 1–12,
Valencia, Spain, April 3, 2017.

tions and domains of raw text, making this the first study ever to address the impact of text understanding perspective on sentence-level emotion annotation. The results we achieved directly influenced the design and creation of EMOBANK, a novel large-scale gold standard for emotion analysis employing the VAD model for affect representation (Buechel and Hahn, 2017).

2 Related Work

Representation Schemes for Emotion. Due to the multi-disciplinary nature of research on emotions, different representation schemes and models have emerged hampering comparison across different approaches (Buechel and Hahn, 2016).

In NLP-oriented sentiment and emotion analysis, the most popular representation scheme is based on *semantic polarity*, the positiveness or negativeness of a word or a sentence, while slightly more sophisticated schemes include a neutral class or even rely on a multi-point polarity scale (Pang and Lee, 2008).

Despite their popularity, these bi- or tri-polar schemes have only loose connections to emotion models currently prevailing in psychology (Sander and Scherer, 2009). From an NLP point of view, those can be broadly subdivided into *categorical* and *dimensional* models (Calvo and Mac Kim, 2013). Categorical models assume a small number of distinct emotional classes (such as *Anger*, *Fear* or *Joy*) that all human beings are supposed to share. In NLP, the most popular of those models are the six *Basic Emotions* by Ekman (1992) or the 8-category scheme of the *Wheel of Emotion* by Plutchik (1980).

Dimensional models, on the other hand, are centered around the notion of compositionality. They assume that emotional states can be best described as a combination of several fundamental factors, i.e., emotional *dimensions*. One of the most popular dimensional models is the Valence-Arousal-Dominance (VAD; Bradley and Lang (1994)) model which postulates three orthogonal dimensions, namely *Valence* (corresponding to the concept of polarity), *Arousal* (a calm-excited scale) and *Dominance* (perceived degree of control in a (social) situation); see Figure 1 for an illustration. An even more wide-spread version of this model uses only the Valence and Arousal dimension, the VA model (Russell, 1980).

For a long time, categorical models were predominant in emotion analysis (Ovesdotter Alm et al., 2005; Strapparava and Mihalcea, 2007; Balahur et al., 2012). Only recently, the VA(D) model found increasing recognition (Paltoglou et al., 2013; Yu et al., 2015; Buechel and Hahn, 2016; Wang et al., 2016). When one of these dimensional models is selected, the task of emotion analysis is most often interpreted as a regression problem (predicting real-valued scores for each of the dimension) so that another set of metrics must be taken into account than those typically applied in NLP (see Section 3).

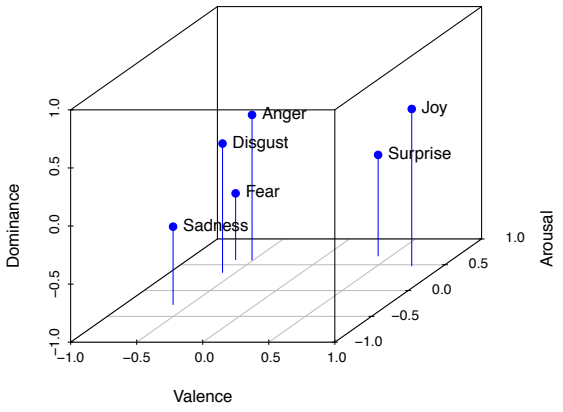

Figure 1: The emotional space spanned by the Valence-Arousal-Dominance model. For illustration, the position of Ekman's six *Basic Emotions* are included (as determined by Russell and Mehrabian (1977)).

Despite its growing popularity, the first large-scale gold standard for dimensional models has only very recently been developed as a follow-up to this contribution (EMOBANK; Buechel and Hahn (2017)). The results we obtained here were crucial for the design of EMOBANK regarding the choice of annotation perspective and the domain the raw data were taken from. However, our results are not only applicable to VA(D) but also to semantic polarity (as Valence is equivalent to this representation format) and may probably generalize over other models of emotion, as well.

Resources and Annotation Methods. For the VAD model, the Self-Assessment Manikin (SAM; Bradley and Lang (1994)) is the most important and to our knowledge only standardized instrument for acquiring emotion ratings based on human self-perception in behavioral psychology (Sander and Scherer, 2009). SAM iconically displays differences in Valence, Arousal and Dominance by a set of anthropomorphic cartoons on

a multi-point scale (see Figure 2). Subjects refer to one of these figures per VAD dimension to rate their feelings as a response to a stimulus.

SAM and derivatives therefrom have been used for annotating a wide range of resources for word-emotion associations in psychology (such as Warriner et al. (2013), Stadthagen-Gonzalez et al. (2016), Yao et al. (2016) and Schmidtke et al. (2014)), as well as VAD-annotated corpora in NLP; Preoţiuc-Pietro et al. (2016) developed a corpus of 2,895 English Facebook posts (but they rely on only two annotators). Yu et al. (2016) generated a corpus of 2,009 Chinese sentences from different genres of online text.

A possible alternative to SAM is Best-Worst Scaling (BSW; Louviere et al. (2015)), a method only recently introduced into NLP by Kiritchenko and Mohammad (2016). This annotation method exploits the fact that humans are typically more consistent when *comparing* two items relative to each other with respect to a given scale rather than *attributing numerical ratings* to the items directly. For example, deciding whether one sentence is more positive than the other is easier than scoring them (say) as 8 and 6 on a 9-point scale.

Although BWS provided promising results for polarity (Kiritchenko and Mohammad, 2016), in this paper, we will use SAM scales. First, with this decision, there are way more studies to compare our results with and, second, the adequacy of BWS for emotional dimensions other than Valence (polarity) remains to be shown.

Perspectival Understanding of Emotions. As stated above, research on the linkage of different annotation perspectives (typically reader *vs.* writer) is really rare. Tang and Chen (2012) examine the relation between the sentiment of microblog posts and the sentiment of their comments (as a proxy for reader emotion) using a positive-negative scheme. They examine which linguistic features are predictive for certain emotion transitions (combinations of an initial *writer* and a responsive *reader* emotion). Liu et al. (2013) model the emotion of a news reader jointly with the emotion of a comment writer using a co-training approach. This contribution was followed up by Li et al. (2016) who criticized that important assumptions underlying co-training, *viz.* sufficiency and independence of the two views, had actually been violated in that work. Instead, they propose a two-view label propagation approach.

Various (knowledge) representation formalisms have been suggested for inferring sentiment or opinions by either readers, writers or both from a piece of text. Reschke and Anand (2011) propose the concept of predicate-specific *evaluativity functors* which allow for inferring the writers' evaluation of a proposition based on the evaluation of the arguments of the predicate. Using description logics as modeling language Klenner (2016) advocates the concept of *polarity frames* to capture polarity constraints verbs impose on their complements as well as polarity implications they project on them. Deng and Wiebe (2015) employ probabilistic soft logic for entity and event-based opinion inference from the viewpoint of the author or intra-textual entities. Rashkin et al. (2016) introduce *connotation frames* of (verb) predicates as a comprehensive formalism for modeling various evaluative relationships (being positive, negative or neutral) between the arguments of the predicate as well as the reader's and author's view on them. However, up until know, the power of this formalism is still restricted by assuming that author and reader evaluate the arguments in the same way.

In summary, different from our contribution, this line of work tends to focus less on the reader's perspective and also addresses cognitive evaluations (*opinions*) rather than instantaneous affective reactions. Although these two concepts are closely related, they are yet different and in fact their relationship has been the subject of a long lasting and still unresolved debate in psychology (Davidson et al., 2003) (e.g., are we afraid of something because we evaluate it as dangerous, or do we evaluate something as dangerous because we are afraid?).

To the best of our knowledge, only Mohammad and Turney (2013) investigated the effects of different perspectives on annotation quality. They conducted an experiment on how to formulate the emotion annotation question and found that asking whether a term is *associated* with an emotion actually resulted in higher IAA than asking whether a term *evokes* a certain emotion. Arguably, the former phrasing is rather unrelated to either writer or reader emotion, while the latter clearly targets the emotion of the reader. Their work renders evidence for the importance of the *perspective* of text comprehension for annotation quality. Note that they focused on word emotion rather than sentence emotion.

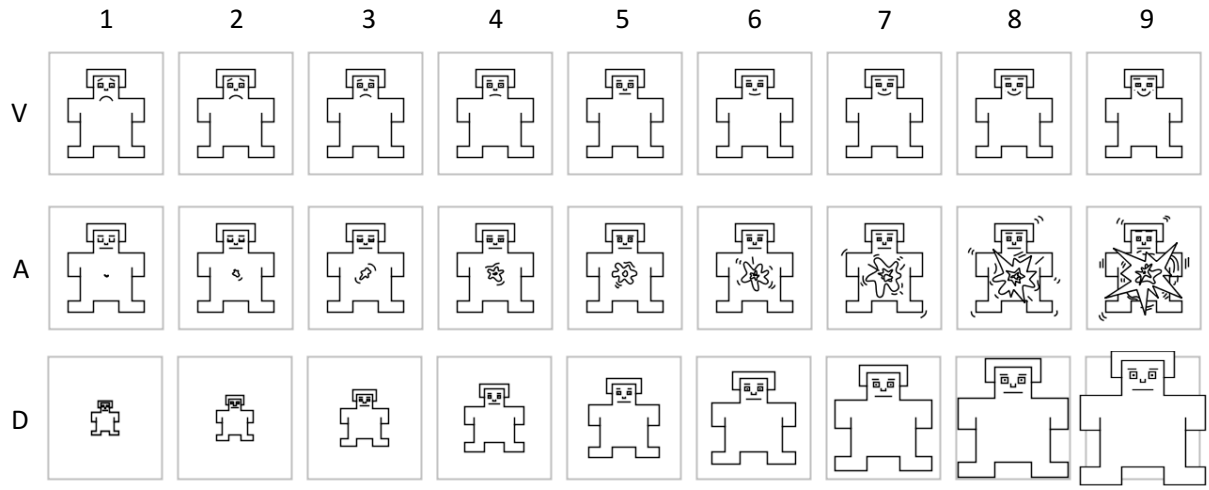

Figure 2: The icons of the 9-point Self-Assessment Manikin (SAM). Dimensions (Valence, Arousal and Dominance; VAD) in rows, rating scores (1-9) in columns. Comprised in PXLab, an open source toolkit for psychological experiments (`http://irtel.uni-mannheim.de/pxlab/index.html`).

3 Methods

Inter-Annotator Agreement. Annotating emotion on numerical scales demands for another statistical tool set than the one that is common in NLP. Well-known metrics such as the κ-coefficient should not be applied for measuring IAA because these are designed for nominal-scaled variables, i.e., ones whose possible values do not have any intrinsic order (such as part-of-speech tags as compared to (say) a multi-point sentiment scale).

In the literature, there is no consensus on what metrics for IAA should be used instead. However, there is a set of repetitively used approaches which are typically only described verbally. In the following, we offer comprehensive formal definitions and a discussion of them.

First, we describe a leave-one-out framework for IAA where the ratings of an individual annotator are compared against the average of the remaining ratings. As one of the first papers, it was used and verbally described by Strapparava and Mihalcea (2007) and was later taken on by Yu et al. (2016) and Preoţiuc-Pietro et al. (2016).

Let $X := (x_{ij}) \in \mathbb{R}^{m \times n}$ be a matrix where m corresponds to the number of items and n corresponds to the number of annotators. X stores all the individual ratings of the m items (organized in rows) and n annotators (organized in columns) so that x_{ij} represents the rating of the i-th item by the j-th annotator. Since we use the three-dimensional VAD model, in practice, we will have one such matrix for each VAD dimension.

Let b_j denote $(x_{1j}, x_{2j}, .., x_{mj})$, the vector composed out of the j-th column of the matrix and let $f : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ be an arbitrary metric for comparing two data series, then $L1O_f(X)$, the leave-one-out IAA for the rating matrix X relative to the metric f, is defined as

$$L1O_f(X) := \frac{1}{n} \sum_{j=1}^{n} f(b_j, b_j^{\varnothing}) \tag{1}$$

where $b_j^{\varnothing}$ is the average annotation vector of the remaining raters:

$$b_j^{\varnothing} := \frac{1}{n-1} \sum_{k \in \{1,..,n\} \setminus \{j\}} b_k \tag{2}$$

For our experiments, we will use three different metrics specifying the function f, namely r, MAE and RMSE.

In general, the Pearson correlation coefficient r captures the linear dependence between two data series, $\mathbf{x} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$ and $\mathbf{y} = \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m$. In our case $\mathbf{x}$,$\mathbf{y}$ correspond to the rating vector of an individual annotator and the aggregated rating vector of the remaining annotators, respectively.

$$r(\mathbf{x}, \mathbf{y}) := \frac{\sum_{i=1}^{m} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})}{\sqrt{\sum_{i=1}^{m} (\mathbf{x}_i - \overline{\mathbf{x}})^2} \sqrt{\sum_{i=1}^{m} (\mathbf{y}_i - \overline{\mathbf{y}})^2}} \tag{3}$$

where $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ denote the mean value of $\mathbf{x}$, $\mathbf{y}$, respectively.

When comparing a model's prediction to the actual data, it can be very important not only to

take correlation-based metrics like r into account, but also error-based metrics (Buechel and Hahn, 2016). This is so because a model may produce very accurate predictions in terms of correlation, while at the same time it may perform poorly when taking errors into account (for instance, when the predicted values range in a much smaller interval than the actual values).

To be able to compare a system's performance more directly to the human ceiling, we also apply error-based metrics within this leave-one-out framework. The most popular ones for emotion analysis are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) (Paltoglou et al., 2013; Yu et al., 2016; Wang et al., 2016):

$$\text{MAE}(\mathbf{x}, \mathbf{y}) := \frac{1}{m} \sum_{i=1}^{m} |(\mathbf{x}_i - \mathbf{y}_i)| \quad (4)$$

$$\text{RMSE}(\mathbf{x}, \mathbf{y}) := \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \mathbf{y}_i)^2} \quad (5)$$

One of the drawbacks of this framework is that each x_{ij} from matrix X has to be known in order to calculate the IAA. An alternative method was verbally described by Buechel and Hahn (2016) which can be computed out of mean and SD values for each item alone (a format often available from psychological papers). Let X be defined as above and let $\overline{a_i}$ denote the mean value for the i-th item. Then, the Average Annotation Standard Deviation (AASD) is defined as

$$\text{AASD}(X) := \frac{1}{m} \sum_{i=1}^{m} \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_{ij} - \overline{a_i})^2} \quad (6)$$

Emotionality. While IAA is indubitably the most important quality criterion for emotion annotation, we argue that there is at least one additional criterion that is not covered by prior research: When using numerical scales (especially ones with a large number of rating points, e.g., the 9-point scales we will use in our experiments) annotations where only neutral ratings are used will be unfavorable for future applications (e.g., training models). Therefore, it is important that the annotations are properly distributed over the full range of the scale. This issue is especially relevant in our setting as different perspectives may very well differ in the extremity of their reactions, as evident from Example (1). We call this desirable property the *emotionality* (EMO) of the annotations.

For the EMO metric, we first derive aggregated ratings from the individual rating decisions of the annotators, i.e., the ratings that would later form the final ratings of a corpus. For that, we aggregate the rating matrix X from Equation 1 into the vector y consisting of the respective row means $\overline{y_i}$.

$$\overline{y_i} := \frac{1}{n} \sum_{j=1}^{n} x_{ij} \quad (7)$$

$$y := (\overline{y_1}, ..., \overline{y_i}, ..., \overline{y_m}) \quad (8)$$

Since we use the VAD model, we will have one such aggregated vector per VAD dimension. We denote them y^1, y^2 and y^3. Let the matrix $Y = (y_i^j) \in \mathbb{R}^{m \times 3}$ hold the aggregated ratings of item i for dimension j, and let $\mathcal{N}$ denote the neutral rating (e.g., 5 on a 9-point scale). Then,

$$\text{EMO}(Y) := \frac{1}{3 \times m} \sum_{j=1}^{3} \sum_{i=1}^{m} |y_i^j - \mathcal{N}| \quad (9)$$

Representative Distribution. A closely related quality indicator relates to the representativeness of the resulting rating distribution. For large sets of stimuli (words as well as sentences), numerous studies consistently report that when using SAM-like scales, typically the emotion ratings closely resemble a normal distribution, i.e., the density plot displays a Gaussian, "bell-shaped" curve (see Figure 3b) (Preoţiuc-Pietro et al., 2016; Warriner et al., 2013; Stadthagen-Gonzalez et al., 2016; Montefinese et al., 2014).

Intuitively, it makes sense that most of the sentences under annotation should be rather neutral, while only few of them carry extreme emotions. Therefore, we argue that ideally the resulting aggregated ratings for an emotion annotation task should be normally distributed. Otherwise, it must be seriously called into question in how far the respective data set can be considered representative, possibly reducing the performance of models trained thereon. Consequently, we will also take the density plot of the ratings into account when comparing different set-ups.

4 Experiments

Perspectives to Distinguish. Considering Example (1) and our literature review from Section

2, it is obvious that at least the perspective of the *writer* and the *reader* of an utterance must be distinguished. Accordingly, writer emotion refers to how someone feels while producing an utterance, whereas reader emotion relates to how someone feels right after reading or hearing this utterance.

Also taking into account the finding by Mohammad and Turney (2013) that agreement among annotators is higher when asking whether a word is *associated* with an emotion rather than asking whether it *evokes* this emotion, we propose to extend the common writer-reader framework by a third category, the *text* perspective, where no actual person is specified as perceiving an emotion. Rather, we assume for this perspective that emotion is an intrinsic property of a sentence (or an alternative linguistic unit like a phrase or the entire text). In the following, we will use the terms WRITER, TEXT and READER to concisely refer to the respective perspectives.

Data Sets. We collected two data sets, a movie review data set highly popular in sentiment analysis and a balanced corpus of general English. In this way, we can estimate the annotation quality resulting from different perspectives, also covering interactions regarding different domains.

The first data set builds upon the corpus originally introduced by Pang and Lee (2005). It consists of about 10k snippets from movie reviews by professional critics collected from the website `rottentomatoes.com`. The data was further enriched by Socher et al. (2013) who annotated individual nodes in the constituency parse trees according to a 5-point polarity scale, forming the Stanford Sentiment Treebank (SST) which contains 11,855 sentences.

Upon closer inspection, we noticed that the SST data have some encoding issues (e.g., *Absorbing character study by AndrÃ© Turpin* .) that are not present in the original Rotten Tomatoes data set. So we decided to replicate the creation of the SST data from the original snippets. Furthermore, we filtered out fragmentary sentences automatically (e.g., beginning with comma, dashes, lower case, etc.) as well as manually excluded grammatically incomplete and therefore incomprehensible sentences, e.g., "Or a profit" or "Over age 15?". Subsequently, a total of 10,987 sentences could be mapped back to SST IDs forming the basis for our experiments (the SST* collection).

To complement our review language data set, a domain heavily focused on in sentiment analysis (Liu, 2015), for our second data set, we decided to rely on a genre-balanced corpus. We chose the Manually Annotated Sub-Corpus (MASC) of the American National Corpus which is already annotated for various linguistic levels (Ide et al., 2008; Ide et al., 2010). We excluded registers containing spoken, mainly dialogic or non-standard language, e.g., telephone conversations, movie scripts and tweets. To further enrich this collection of raw data for potential emotion analysis applications, we additionally included the corpus of the SEMEVAL-2007 Task 14 focusing on *Affective Text* (SE07; Strapparava and Mihalcea (2007)), one of the most important data sets in emotion analysis. This data set already bears annotations according to Ekman's six Basic Emotions (see Section 2) so that the gold standard we ultimately supply already contains a bi-representational part (being annotated according to a dimensional *and* a categorical model of emotion). Such a double encoding will easily allow for research on automatically mapping between different emotion formats (Buechel and Hahn, 2017).

In order to identify individual sentence in MASC, we relied on the already available annotations. We noticed, however, that a considerable portion of the sentence boundary annotations were duplicates which we consequently removed (about 5% of the preselected data). This left us with a total of 18,290 sentences from MASC and 1,250 headlines from SE07. Together, they form our second data set, MASC*.

Study Design. We pulled a 40 sentences random sample from MASC* and SST*, respectively. For each of the three perspectives WRITER, READER and TEXT, we prepared a separate set of instructions. Those instructions are identical, except for the exact phrasing of what a participant should annotate: For WRITER, it was consistently asked "what emotion is expressed by the author", while TEXT and READER queried "what emotion is conveyed" by and "how do you [the participant of the survey] feel after reading" an individual sentence, respectively.

After reviewing numerous studies from NLP and psychology that had created emotion annotations (e.g., Katz et al. (2007), Strapparava and Mihalcea (2007), Mohammad and Turney (2013), Pinheiro et al. (2016), Warriner et al. (2013)), we largely relied on the instructions used by Bradley

and Lang (1999) as this is one of the first and probably the most influential resource from psychology which also greatly influenced work in NLP (Yu et al., 2016; Preoţiuc-Pietro et al., 2016).

The instructions were structured as follows. After a general description of the study, the individual scales of SAM were explained to the participants. After that, they performed three trial ratings to familiarize themselves with the usage of the SAM scales before proceeding to judge the actual 40 sentences of interest. The study was implemented as a web survey using Google Forms.[1] The sentences were presented in randomized order, i.e., they were shuffled for each participant individually.

For each of the six resulting surveys (one for each combination of perspective and data set), we recruited 80 participants via the crowdsourcing platform `crowdflower.com` (CF). The number was chosen so that the differences in IAA may reach statistical significance (according to the leave-one-out evaluation (see Section 3), the number of cases is equal to the number of raters). The surveys went online one after the other, so that as few participants as possible would do more than one of the surveys. The task was available from within the UK, the US, Ireland, Canada, Australia and New Zealand.

We preferred using an external survey over running the task directly via the CF platform because this set-up offers more design options, such as randomization, which is impossible via CF; there, the data is only shuffled once and will then be presented in the same order to each participant. The drawback of this approach is that we cannot rely on CF's quality control mechanisms.

In order to still be able to exclude malicious raters, we introduced an algorithmic filtering process where we summed up the absolute error the participants made on the trial questions—those were asking them to indicate the VAD values for a verbally described emotion so that the correct answers were evident from the instructions. Raters whose absolute error was above a certain threshold were excluded.

We set this parameter to 20 (removing about a third of the responses) because this was approximately the ratio of raters which struck us as unreliable when manually inspecting the data while, at the same time, leaving us with a reasonable number of cases to perform statistical analysis. The results of this analysis is presented in the following section. Our two small sized yet multi-perspectival data sets are publicly available for further analysis.[2]

[1] `https://forms.google.com/`

	Perspective	r	MAE	RMSE	AASD
SST*	WRITER	.53	1.41	1.70	1.73
	TEXT	.41	1.73	2.03	2.10
	READER	.40	1.66	1.96	2.02
MASC*	WRITER	.43	1.56	1.88	1.95
	TEXT	.43	1.49	1.81	1.89
	READER	.36	1.58	1.89	1.98

Table 1: IAA values obtained on the SST* and the MASC* data set. r, MAE and RMSE refer to the respective leave-one-out metric (see Section 3).

5 Results

In this section, we compare the three annotation perspectives (WRITER, READER and TEXT) on two different data sets (SST* and MASC*; see Section 4), according to three criteria for annotation quality: IAA, emotionality and distribution (see Section 3).

Inter-Annotator Agreement. Since there is no consensus on a fixed set of metrics for numerical emotion values, we compare IAA according to a range of measures. We use r, MAE and RMSE in the leave-one-out framework, as well as AASD (see Section 3). Table 1 displays our results for the SST* and MASC* data set. We calculated IAA individually for Valence, Arousal and Dominance. However, to keep the number of comparisons feasible, we restrict ourselves to presenting the respective mean values (average over VAD), only. The relative ordering between the VAD dimensions is overall consistent with prior work so that Valence shows better IAA than Arousal or Dominance (in line with findings from Warriner et al. (2013) and Schmidtke et al. (2014)).

We find that on the review-style SST* data, WRITER displays the best IAA according to all of the four metrics ($p < 0.05$ using a two-tailed t-test, respectively). Note that MAE, RMSE and AASD are error-based so that the smaller the value the better the agreement. Concerning the ordering of the remaining perspectives, TEXT is marginally better regarding r, while the results from the three error-based metrics are clearly in favor of READER. Consequently, for IAA on the

[2] `https://github.com/JULIELab/EmoBank`

	Perspective	EMO
SST*	WRITER	1.09
	TEXT	1.04
	READER	0.91
MASC*	WRITER	0.75
	TEXT	0.70
	READER	0.63

Table 2: Emotionality results for the SST* and the MASC* data set.

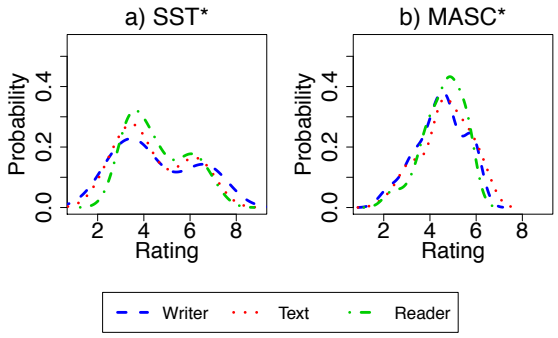

Figure 3: Density plots of the aggregated Valence ratings for the two data sets and three perspectives.

SST* data set, WRITER yields the best performance, while the order of the other perspectives is not so clear.

Surprisingly, the results look markedly different on the MASC* data. Here, regarding r, WRITER and TEXT are on par with each other. This contrasts with the results from the error-based metrics. There, TEXT shows the best value, while WRITER, in turn, improves upon READER only by a small margin. Most importantly, for neither of the four metrics we obtain statistical significance between the best and the second best perspective ($p \geq 0.05$ using a two-tailed t-test, respectively). Thus, concerning IAA on the MASC* sample, the results remain rather opaque.

The fact that, contrary to that, on SST* the results are conclusive and statistically significant, strongly suggests that the resulting annotation quality is not only dependent on the annotation perspective. Instead, there seem to be considerable dependencies and interactions concerning the domain of the raw data, as well.

Interestingly, on both corpora correlation- and error-based sets of metrics behave inconsistently which we interpret as a piece of evidence for using both types of metrics, in parallel (Buechel and Hahn, 2016; Wang et al., 2016).

Emotionality. For emotionality, we rely on the EMO metric which we defined in Section 3 (see Table 2 for our results). For both corpora, the ordering of the perspectives according to the EMO score is consistent: WRITER yields the most emotional ratings followed by TEXT and READER. ($p < 0.05$ for each of the pairs using a two-tailed t-test). These unanimous and statistically significant results further underpin the advantage of the TEXT and especially the WRITER perspective as already suggested by our findings for IAA.

Distribution. We also looked at the distribution of the resulting aggregated annotations relative to the chosen data sets and the three perspectives by examining the respective density plots. In Figure 3, we give six examples of these plots, displaying the Valence density curve for both corpora, SST* and MASC*, as well as the three perspectives. For Arousal and Dominance, the plots show the same characteristics although slightly less pronounced.

The left density plots, for the SST*, display a bimodal distribution (having two local maxima), whereas the MASC* plots are much closer to a normal distribution. This second shape has been consistently reported by many contributions (see Section 3), whereas we know of no other study reporting a bimodal emotion distribution. This highly atypical finding for SST* might be an artifact of the website from which the original movie review snippets were collected—there, movies are classified into either *fresh* (positive) or *rotten* (negative). Consequently, this binary classification scheme might have influenced the selection of snippets from full-scale reviews (as performed by the website) so that these snippets are either clearly positive or negative.

Thus, our findings seriously call into question in how far the movie review corpus by Pang and Lee (2005)—one of the most popular data sets in sentiment analysis—can be considered representative for review language or general English. Ultimately, this may result in a reduced performance of models trained on such skewed data.

6 Discussion

Overall, we interpret our data as suggesting the WRITER perspective to be superior to TEXT and READER: Considering IAA, it is significantly better on one data set (SST*), while it is on par with or only marginally worse than the best perspective on the other data set (MASC*). Regarding emotionality of the aggregated ratings (EMO), the superiority of this perspective is even more obvious.

The relative order of TEXT and WRITER on the other hand, is not so clear. Regarding IAA, TEXT is better on MASC* while for SST* READER seems to be slightly better (almost on par regarding r but markedly better relative to the error measures we propose here). However, regarding the emotionality of the ratings, TEXT clearly surpasses READER.

Our data suggest that the results of Mohammad and Turney (2013) (the only comparable study so far, though considering emotion on the *word* rather than *sentence* level) may be also true for sentences in most of the cases. However, our data indicate that the validity of their findings may depend on the domain the raw data originate from. They found that phrasing the emotion annotation task relative to the TEXT perspective yields higher IAA than relating to the READER perspective. However, more importantly, our data complement their results by presenting evidence that WRITER seems to be even better than any of the two perspectives they took into account.

7 Conclusion

This contribution presented a series of annotation experiments examining which *annotation perspective* (WRITER, TEXT or READER) yields the best IAA, also taking domain differences into account—the first study of this kind for sentence-level emotion annotation. We began by reviewing different popular representation schemes for emotion before (formally) defining various metrics for annotation quality—for the VAD scheme we use, this task was so far neglected in the literature.

Our findings strongly suggest that WRITER is overall the superior perspective. However, the exact ordering of the perspectives strongly depends on the domain the data originate from. Our results are thus mainly consistent with, but substantially go beyond, the only comparable study so far (Mohammad and Turney, 2013). Furthermore, our data provide strong evidence that the movie review corpus by Pang and Lee (2005)—one of the most popular ones for sentiment analysis—may not be representative in terms of its rating distribution potentially casting doubt on the quality of models trained on this data.

For the subsequent creation of EMOBANK, a large-scale VAD gold standard, we took the following decisions in the light of these not fully conclusive outcomes. First, we decided to annotate a 10k sentences subset of the MASC* corpus considering the atypical rating distribution in the SST* data set. Furthermore, we decided to annotate the whole corpus bi-perspectivally (according to WRITER *and* READER viewpoint) as we hope that the resulting resource helps clarifying which factors exactly influence emotion annotation quality. This freely available resource is further described in Buechel and Hahn (2017).

References

A. Balahur, J. M. Hermida, and A. Montoyo. 2012. Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1):88–101.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In Gal A. Kaminka, Maria Fox, Paolo Bouquet, Eyke Hüllermeier, Virginia Dignum, Frank Dignum, and Frank van Harmelen, editors, *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. The Hague, The Netherlands, August 29 - September 2, 2016*, pages 1114–1122.

Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Annual Meeting of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, April 3-7, 2017*.

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith. 2003. *Handbook of Affective Sciences*. Oxford University Press, Oxford, New York, NY.

Lingjia Deng and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.*

Lisbon, Portugal, September 17–21, 2015, pages 179–189.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Nancy C. Ide, Collin F. Baker, Christiane Fellbaum, Charles J. Fillmore, and Rebecca J. Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan E. J. M. Odijk, Stelios Piperidis, and Daniel Tapias, editors, *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, May 26 – June 1, 2008*, pages 2455–2461.

Nancy C. Ide, Collin F. Baker, Christiane Fellbaum, and Rebecca J. Passonneau. 2010. The Manually Annotated Sub-Corpus: A community resource for and by the people. In Jan Hajič, M. Sandra Carberry, and Stephen Clark, editors, *ACL 2010 — Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, July 11–16, 2010*, volume 2: Short Papers, pages 68–73.

Phil Katz, Matthew Singleton, and Richard Wicentowski. 2007. SWAT-MP: The SemEval-2007 systems for Task 5 and Task 14. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *SemEval-2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007. Prague, Czech Republic, June 23-24, 2007*, pages 308–313.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 811–817.

Manfred Klenner. 2016. A model for multi-perspective opinion inferences. In Larry Birnbaum, Octavian Popescu, and Carlo Strapparava, editors, *Proceedings of IJCAI 2016 Workshop Natural Language Meets Journalism, New York, USA, July 10, 2016*, pages 6–11.

Shoushan Li, Jian Xu, Dong Zhang, and Guodong Zhou. 2016. Two-view label propagation to semi-supervised reader emotion classification. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan, December 11-16, 2016*, volume Technical Papers, pages 2647–2655.

Hsin-Yih Kevin Lin and Hsin-Hsi Chen. 2008. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *EMNLP 2008 — Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, October 25–27, 2008*, pages 136–144.

Huanhuan Liu, Shoushan Li, Guodong Zhou, Chu-Ren Huang, and Peifeng Li. 2013. Joint modeling of news reader's and comment writer's emotions. In Hinrich Schütze, Pascale Fung, and Massimo Poesio, editors, *ACL 2013 — Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, August 4-9, 2013*, volume 2: Short Papers, pages 511–515.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, New York.

Jordan J Louviere, Terry N Flynn, and AAJ Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, Cambridge.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46(3):887–903.

Subhabrata Mukherjee and Sachindra Joshi. 2014. Author-specific sentiment aggregation for polarity prediction of reviews. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Loftsson Hrafn, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 3092—3099.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In Raymond J. Mooney, Christopher Brew, Lee-Feng Chien, and Katrin Kirchhoff, editors, *HLT-EMNLP 2005 — Proceedings of the Human Language Technology Conference & 2005 Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, 6-8 October 2005*, pages 579–586.

G. Paltoglou, M. Theunis, A. Kappas, and M. Thelwall. 2013. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 4(1):106–115.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Kevin Knight, Tou Hwee Ng, and Kemal Oflazer, editors, *ACL 2005 — Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann*

Arbor, Michigan, USA, June 25–30, 2005, pages 115–124.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Ana P. Pinheiro, Marcelo Dias, Joo Pedrosa, and Ana P. Soares. 2016. Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*. Online First Publication.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research and Experience*, 1(3):3–33.

Daniel Preoţiuc-Pietro, Hansen Andrew Schwartz, Gregory Park, Johannes C. Eichstaedt, Margaret L. Kern, Lyle H. Ungar, and Elizabeth P. Shulman. 2016. Modelling valence and arousal in Facebook posts. In Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andrés Montoyo, editors, *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, pages 9–15.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In Antal van den Bosch, Katrin Erk, and Noah A. Smith, editors, *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7–12, 2016*, volume 1: Long Papers, pages 311–321.

Kevin Reschke and Pranav Anand. 2011. Extracting contextual evaluativity. In Johan Bos and Stephen Pulman, editors, *IWCS 2011 — Proceedings of the 9th International Conference on Computational Semantics. Oxford, UK, January 12–14, 2011*, pages 370–374.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

David Sander and Klaus R. Scherer, editors. 2009. *The Oxford Companion to Emotion and the Affective Sciences*. Oxford University Press, Oxford, New York.

David S. Schmidtke, Tobias Schröder, Arthur M. Jacobs, and Markus Conrad. 2014. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, 46(4):1108–1118.

Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Timothy Baldwin and Anna Korhonen, editors, *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA, 18-21 October 2013*, pages 1631–1642.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez Sánchez, and Marc Brysbaert. 2016. Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*. Online First Publication.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *SemEval-2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007. Prague, Czech Republic, June 23-24, 2007*, pages 70–74.

Yi-jie Tang and Hsin-Hsi Chen. 2012. Mining sentiment words from microblogs for predicting writer-reader emotion transition. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan E. J. M. Odijk, and Stelios Piperidis, editors, *LREC 2012 — Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012*, pages 1226–1229.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In Antal van den Bosch, Katrin Erk, and Noah A. Smith, editors, *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7–12, 2016*, volume 2: Short Papers, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbært. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhendong Wang. 2016. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*. Online First Publication.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In Yuji Matsumoto, Chengqing Zong, and Michael Strube, editors, *ACL-IJCNLP 2015 — Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics & 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, July 26–31, 2015*, volume 2: Short Papers, pages 788–793.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In Kevin C. Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12–17, 2016*, pages 540–545.