ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1114

Emotion Analysis as a Regression Problem — Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation

Sven Buechel and **Udo Hahn**[1]

Abstract. Emotion analysis (EA) and sentiment analysis are closely related tasks differing in the psychological phenomenon they aim to catch. We address fine-grained models for EA which treat the computation of the emotional status of narrative documents as a regression rather than a classification problem, as performed by coarse-grained approaches. We introduce Ekman's Basic Emotions (BE) and Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model—two major schemes of emotion representation following opposing lines of psychological research, i.e., categorical and dimensional models—and discuss problems when BEs are used in a regression approach. We present the first natural language system thoroughly evaluated for fine-grained emotion analysis using the VAD scheme. Although we only employ simple BOW features, we reach correlation values up until $r = .65$ with human annotations. Furthermore, we show that the prevailing evaluation methodology relying solely on Pearson's correlation coefficient r is deficient which leads us to the introduction of a complementary error-based metric. Due to the lack of comparable (VAD-based) systems, we, finally, introduce a novel method of mapping between VAD and BE emotion representations to create a reasonable basis for comparison. This enables us to evaluate VAD output against human BE judgments and, thus, allows for a more direct comparison with existing BE-based emotion analysis systems. Even with this, admittedly, error-prone transformation step our VAD-based system achieves state-of-the-art performance in three out of six emotion categories, out-performing all existing BE-based systems but one.

1 Introduction

Affective states expressed via written or spoken utterances, as well as non-verbal gestures and mimics in discourse are at the core of any cognitively plausible theory of human communication. From a computational perspective, AI researchers have already started investigating into this field [26], since progress in this area will pave the way to even smarter and more natural computational agents for human-computer interaction, such as avatars or robots .

However, this research area at the intersection of (cognitive) psychology, (computational) linguistics, and artificial intelligence suffers from some confusing uses of terminology [22] which have to be sorted out before we get started. Following Pang and Lee [25] we subsume all work done in this area under the umbrella term *subjectivity analysis*. Its most widespread subtask is sentiment analysis or opinion mining (both terms are used interchangeably). In this work, we address another subtask which has recently become more and more popular, namely *emotion analysis* (EA). From a representational perspective, *sentiment* typically refers to the semantic polarity (the positiveness or negativeness relative to some target entity) of a sentence or a document. While sentiment analysis has usually only loose (or no) ties to models taken from psychology, *emotion* (describing phenomena such as anger, fear, or joy) is often represented in a more complex way making direct use of larger pieces of psychological theory.

There are two main dividing lines in the field of EA. The first one (as discussed, e.g., by Calvo and Kim [10]) relates to the choice of a psychological model. Following *categorical models*, emotional states can be subcategorized into a small set of emotion categories. Ekman's Basic Emotion (BE) model [14] is perhaps the most influential among those categorical approaches. On the other hand, following *dimensional models* an emotional state is described relative to a small number of *emotional dimensions*. Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model [28] is among the most commonly used dimensional approaches.

The second and maybe even more fundamental dividing line (as discussed, e.g., by Strapparava and Mihalcea [34]) relates to the main type of predictive problem one faces here. Most of the previous work on EA is *coarse-grained* in the sense that the task of predicting emotion is phrased as a *classification* problem—the output of a corresponding system represents an emotional value as one or multiple class labels. In contrast, *fine-grained* EA treats the task of recognizing emotions as a *regression* problem so that (most often) a vector of real-valued numbers will be produced as the result of an emotion assessment. Note that the choices regarding these dividing lines are made independently from one another, e.g., also allowing for a coarse-grained analysis using dimensional models [16].

The coarse-grained approach seems to be particularly appropriate for highly opinionated social media texts (such as blogs, chats or tweets) but is less likely to account for more subtle expressions of emotions as, e.g., in literary documents (mainly studied in the emerging field of digital humanities [1, 38]), public and personal health narratives (mainly studied in the field of biomedical and clinical NLP [13, 30]) or socio-economic texts (newspaper, newswire, formal business reporting notes, etc. which are increasingly dealt with in computational social science and economics [3, 15, 9]).

In this paper, we focus exclusively on fine-grained emotion analysis. We, first, provide a critical comparison of the BE and the VAD emotion model, as well as a complete survey of prior systems for fine-grained EA. We then present the first VAD-based system for

[1] Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany, URL: http://www.julielab.de

fine-grained EA. Evaluating its performance revealed systematic deficiencies in the evaluation methodology for such systems which lead us to propose a complementary metric. In an attempt to compare our dimensional system more directly with already existing categorical ones, we developed a novel method for mapping between VAD and BE representation schemes and, given these (imperfect) mappings, we find evidence that our system is still among the best-performing systems for predicting the emotional status of narratives.

2 Related Work

2.1 Dimensional versus Categorical Models

Researchers in NLP and psychology have devised a multitude of different models of emotion which can be roughly subdivided into categorical and dimensional models [29, 10, 33]. In computational studies, categorical models most often employ Ekman's [14] six basic emotions (BE: anger, disgust, fear, joy, sadness and surprise) or a derivative therefrom. According to this psychological theory, all human beings share a common set of cross-culturally universal (basic) emotions so that each emotional state of an individual can be unambiguously classified as one of these. Dimensional approaches, on the other hand, often refer to Russell and Mehrabian's Valence-Arousal-Dominance (VAD) model [28].[2] According to this model, emotional states can be described relative to three fundamental emotional dimensions: Valence (the degree of pleasure or displeasure of an emotion), Arousal (level of mental activity, ranging from low engagement to ecstasy) and Dominance (extent of control felt in a given situation). Accordingly, emotions are characterized on three dimensions, each of which spans an interval of real-valued numbers indicating the strength and orientation on each dimension. Providing a fine-grained representation using the VAD model (a vector of real-valued numbers) is therefore straightforward. For BE models, this is typically accomplished by assigning an agreement score to each of the basic emotions (e.g., in the interval [0,100] as realized in the SemEval-2007 test corpus for the *Affective Text* task [34]).

To further illustrate the relationship between the VAD and the BE model, Figure 1 depicts the position of Ekman's basic emotions within the emotional space spanned by the Valence, Arousal and Dominance axis of the VAD model. The assessments were empirically determined by requesting several subjects to describe the six basic emotions in terms of these three dimensions [28]. For fine-grained approaches, we consider VAD to be superior to BE due to the following considerations:

- As Figure 1 reveals, the basic emotions are unevenly distributed in the VAD space. While half of them (anger, disgust and fear) are marked by high arousal and low valence (and therefore reside in one quarter of the space), none of them exhibits high valence and low arousal specifying an emotion like calmness or content. Thus, trying to detect such emotions using a BE-based system may encounter serious problems. Exactly these kinds of emotions have been shown to be most beneficial for the prediction of stock market prices in previous work [3].
- Although Ekman's six-category system is most commonly used, there is no consensus on a fixed set of basic emotions, neither in psychology [29], nor in AI (cf., e.g., [23] and [32]). Not only does this hamper comparison across systems but also does it force researchers to choose different sets of emotional categories according to the emotions which they think to be most relevant for a given application (instead of using a generic and universal representation scheme). This may lead to study designs (e.g., [13]) using a total of 15 different categories considered to indicate suicidal tendencies, e.g., hopelessness or sorrow.
- It is intuitively clear that BEs are *not* equidistant, e.g., fear is obviously more similar to disgust than it is to joy—an observation also supported by Figure 1. Therefore (unlike vectorial VAD representations), distances between given emotions in fine-grained BE representation cannot be meaningfully calculated assuming a vector space with orthogonal axis. This property seriously limits the possibility for further analysis of emotion distributions (such as clustering) and may pose problems for the use of emotion values as features in machine learning.

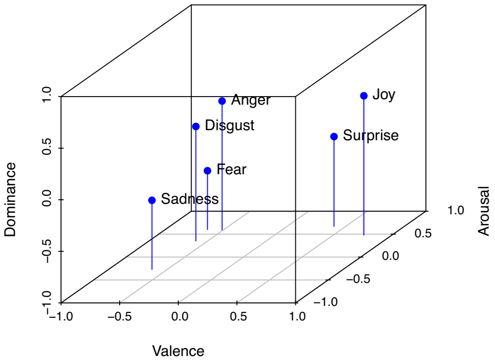

Figure 1. Positions of Ekman's basic emotions within the emotional space spanned by the Valence, Arousal and Dominance axis of the VAD model. Ratings are taken from Russell and Mehrabian [28].

2.2 Computational Resources for Emotion Analysis

In psychology, both models, Ekman's BE as well as Russell and Mehrabian's VAD model, are widely used as standard models [33]. While the VAD model and other dimensional models are commonly preferred in some areas of affective computing [8], NLP researchers, especially those dealing with written documents, almost exclusively subscribe to categorical approaches, most often Ekman's model [10]. As a consequence, these preferences for one model or the other are reflected by the types of resources made available.

Concerning emotion lexicons following the VAD model, the *Affective Norms for English Words* (ANEW) [5] has been most influential in psychological research and was also adapted for many languages other than English [39]. The developers of ANEW asked subjects to rate their feelings on the three VAD dimensions when reading certain words as stimuli. Their responses were encoded using the *Self-Assessment Manikin* (SAM), an icon-style graphical format which consists of three sequences of human-like pictograms, each representing a 9-point scale for Valence, Arousal and Dominance, respectively [4]. The average rating per word was calculated, thus forming its emotional value. The original version of ANEW comprised 1,034 lexical entries. By now, an extended version has been developed amounting to 2,476 words [7].

Bestgen and Vincze [2] extended the original ANEW version by using a bootstrapping method based on Latent Semantic Analysis (LSA) [12]. Their major achievement employing these methods is that they attribute VAD values to formerly unrated words by locating them together with their least distant neighbors whose emotion values are known from the original ANEW resource in a latent semantic

[2] Alternative names for these dimensions include *Pleasure* instead of *Valence* (PAD) as well as *Control* instead *Dominance* (PAC).

space and averaging these values. Re-assessing words already known from Anew, they compute correlations ($r = 0.71$, 0.56 and 0.60 for Valence, Arousal and Dominance, respectively) between the original and the bootstrapped values. Their lexical resource (BV) incorporates 17,350 entries.

Warriner et al. [39] replicated and extended the original Anew lexicon in a crowdscourcing campaign using the *Amazon Mechanical Turk* (AMT). Their resource (WKB) contains more than ten times the entries of Anew (13,915 in total) and excels with particularly high correlations with the original ratings ($r = 0.95$, 0.76 and 0.80 for VAD, respectively). This result is consistent with earlier findings that non-expert ratings for natural language tasks acquired via AMT are, in fact, of good quality (especially when rating emotions) compared to expert ratings [31].

Concerning BE lexicons with fine-grained ratings, Staiano et al. [32] built DepecheMood (DM), a lexical resource which contains more than 37k entries. They exploit the functionality of the social news network `rappler.com` in which users may report their "mood" when reading a piece of news. DepecheMood is constructed by multiplying the document-emotion matrix and the document-term matrix of all available mood-rated articles. The latter was computed using either absolute frequency, normalized frequency and TF-IDF scores, thus leading to three versions of the emotion-term matrix.

Another major resource for tackling emotions is WordNet-Affect (WN-A). It contains both, sentiment assessments (positive, negative, neutral and ambiguous) and a hierarchy of various emotion categories [36, 37]. Though not providing continuous ratings for these categories, previous work on fine-grained analysis has largely relied on this resource (cf. Section 2.4).

Corpora carrying VAD annotations are more than rare. To the best of our knowledge, the *Affective Norms for English Text* (Anet) collection [6] is the only available resource and, up until now, has not been used for NLP tasks. With 120 sentences or short texts, e.g., *"You are lying in bed on a Sunday morning"*, it is truly a tiny little corpus. Its VAD annotations were empirically elicited from subjects using Sam (see above). Most recently, another larger resource (FB) carrying at least Valence and Arousal annotations has been generated [27] which comprises 2,895 Facebook posts rated by two annotators.

Corpora annotated with fine-grained emotion categories are rare, as well. To our knowledge, the corpus provided for the *Affective Text* task of SemEval-2007 [34] is the only one, whereas for coarse-grained annotations, there are much more alternatives; cf. [10, 23]. The SemEval corpus (SE7) contains headlines from major newspapers and consists of two subsets, a development set handed out to the competitors (250 headlines) and a final test set (1,000 headlines). The corpus was independently labeled by six annotators according to the BE model so that an agreement score ranging between $[0, 100]$ could be determined for each headline and emotion. Our survey of computational resources is summarized in Table 1.

Two studies [34, 31] report inter-annotator agreement (IAA) measurements for fine-grained BE labeling (see Table 2). Here, IAA is typically measured, first, by calculating Pearson's correlation between each individual annotator and the average annotation of the other annotators (resulting in one correlation value per rater) and then averaging these values [35]. Additionally, Katz et al. [18] provide the agreement of the overlap of their own annotated corpus and SE7. Both are averages of multiple human annotations and are therefore not comparable to IAA values.

Table 1. Resources for emotion detection (lexicons (Lex) and corpora (Corp)) listing the model of emotion they use, the granularity of ratings (Grain), the acquisition methodology (manual (without further specification), asking subjects in a controlled experimental environment (exp), bootstrapping or crowdsourcing (boot or crowd, respectively) and their size in terms of lexical entries (for lexicons) or sentences/documents (for corpora).

	Acronym	Study	Model	Grain	Method	Size
Lex						
	WN-A	[36, 37]	BE	coarse	manual	1,637
	ANEW	[5]	VAD	fine	exp	1,034
	BV	[2]	VAD	fine	boot	17,350
	WKB	[39]	VAD	fine	crowd	13,915
	DM	[32]	BE	fine	crowd[3]	37,771
Corp						
	ANET	[6]	VAD	fine	exp	120
	SE7	[34]	BE	fine	exp	1,250
	FB	[27]	VA	fine	exp	2,895

Table 2. IAA for fine-grained emotion detection measured in r. From the many IAA values reported by Snow et al. [31], we here include their expert *vs.* expert IAA measurements. For comparison, the average is computed only taking anger, fear, joy and sadness into account.

Study	Anger	Disg.	Fear	Joy	Sadness	Surpr.	**Avg.**
[34]	.496	.445	.638	.599	.682	.361	.604
[31]	.459	.583	.711	.596	.645	.464	.603

The IAA presented in the first two studies—ranging between approximately $r = 0.35$ and 0.70—illustrates the hardness of the task.

In contrast to BE-based corpora, no IAAs are provided for the VAD-based Anet corpus. However, the average standard deviation between ratings for the same instance amounts to $\text{SD} = 1.45$, 1.85 and 1.87 for Valence, Arousal and Dominance, respectively. The fact that the ratings for the latter two are less consistent than for the former one has been observed in a multitude of studies comparing word ratings, as well as whole lexicons [39, 2]. Preoţiuc-Pietro et al. [27] report an IAA on their FB corpus of $r = .768$ and $.827$ for Valence and Arousal, respectively.

2.3 Mappings between Emotion Models

Only few studies deal with the translation between different emotion schemes. Moreover, most of these activities are only concerned with discrete representations of the BE model (i.e., disregarding continuous agreement scores per category). Having a robust, high-accuracy mapping schema for both representations may help further unify both lines of research (in AI, not limited to NLP, as well as in psychology) [33] and would allow for the interchangeable use of resources developed with respect to one model or the other.

In an early study, Russell et al. [28] presented 300 subjects a list of emotion (or feeling) designating words, including terms referring to the basic emotions, and asked them to assess the designated emotions relative to Valence, Arousal and Dominance. The results can thus be used as a simple yardstick for mapping between basic emotions (in discrete representation) and the VAD model (in dimensional representation) as demonstrated in Figure 1. In a similar, much more recent study, Hoffmann et al. [17] asked 70 subjects to position 22 emotion categories (according to the OCC model [24]) in the VAD space via a user-friendly visual tool. They find high inter-subjective consistency between the assessments although variance was markedly higher for Arousal and Dominance.

[3] Rather than directly using crowdsourced word-emotion ratings, DM was calculated using emotionally crowd-annotated newswire material.

Calvo and Kim [10] map VAD values onto a variation of the six basic emotions by computing the position of the emotional categories in the VAD space as the centroid of several keywords (representative for this category) according to the ANEW lexicon. Then, they calculate cosine similarity between an arbitrary VAD emotion and an emotional category and, finally, map these onto another, if the similarity is above a certain threshold, or map it onto *neutral*, otherwise.

Stevenson et al. [33] collect ratings for five of six emotional categories taken from the BE model for the entries of the original ANEW lexicon (so far having only VAD ratings) by questioning 299 subjects. Thus, a multi-model lexicon is created. They perform linear regression and find evidence which suggest *non*-linear dependencies to hold between these two representation schemes, thus hinting at the insufficiency of their predictive models. Note that this is the only study presented here using a continuous representation for input *and* target variables.

2.4 Fine-Grained Emotion Analysis Systems

As already mentioned, in comparison to coarse-grained approaches, fine-grained emotion detection is a rather neglected task. Together with the small amount of annotated text corpora for fine-grained emotion models, we currently face a situation where system development is hampered by the lack of appropriate resources and evaluations deliver only spurious results. Next, we present each system for fine-grained emotion detection we are aware of. For BE systems, the SE7 corpus has been used for evaluation exclusively (although, additionally, other corpora may be used as well when evaluating their performance in coarse-grained settings). The available evaluation results are presented in Table 3. For comparison, the presented average performance takes into account only Anger, Fear, Joy and Sadness, since DM-f does not measure Disgust, whereas our system (see Section 3) fails to compute Surprise (due to limitations of the mapping functions rather than an inherent shortcoming of our system itself).

WNAP [35] is designed as a baseline by computing emotion values directly related to the frequency of WORDNET-AFFECT terms present in a given document. Surprisingly, this very simple keyword-based approach already outperforms three other systems: LSA-ES, LSA-SW and LSA-AEW [35]. Each of these systems uses a *pseudo-document* method by which both, the emotion categories, as well as the individual documents are represented in a semantic space derived from the BNC corpus[4] using LSA. They differ from each other by the words constituting the pseudo-documents which represent an emotion. LSA-SW uses only the word denoting the emotion, LSA-ES adds the whole WORDNET synset, while LSA-AEW uses each synonym of each synset labeled with this emotion according to WORDNET-AFFECT. Obviously, this methods does not seem to be appropriate for the task of fine-grained emotion analysis.

NB-BLOG [35], the only machine learning approach among the BE systems, uses a Naive Bayes classifier. Its performance merely surpasses the baseline. However, it was trained on blog posts rather than news headlines, a shortcoming which may very well account for a great deal of its poor results.

Similarly, the information theory-based UA system [19] shows only slightly better performance than the keyword baseline. It computes the association between a document and an emotion using statistics from Web search engines and measures the proximity between them using pointwise mutual information (PMI). Note that without its apparent difficulty in detecting Joy, the performance would be markedly better.

Table 3. Performance of BE-based systems for fine-grained emotion analysis measured in r. For comparison, the average (Avg) is computed only over Anger, Fear, Joy and Sadness (Sad) (in addition, we report values for Disgust (Dis) and Surprise (Sur)).

System	Anger	Dis	Fear	Joy	Sad	Sur	**Avg**
DM-f	**.360**	—	**.560**	**.390**	**.480**	**.250**	**.448**
AAM	.329	.130	.449	.213	.436	.064	.356
UPAR7	.323	.129	.449	.225	.410	.167	.352
SWAT	.245	**.186**	.325	.261	.390	.118	.305
UA	.232	.162	.232	.024	.123	.078	.152
NB-BLOG	.198	.048	.074	.138	.160	.031	.143
WNAP	.121	-.016	.249	.103	.086	.031	.140
LSA-ES	.178	.074	.181	.063	.133	.121	.139
LSA-SW	.083	.135	.296	.049	.081	.097	.127
LSA-AEW	.058	.083	.103	.070	.107	.124	.084

The upper half of Table 3 is exclusively populated by lexicon-based approaches with or without incorporation of additional linguistic rules for fine-tuning. UPAR7 [11] and AAM [23] both revise lexicon-based word ratings using syntax-oriented rules. The former system boosts the importance of certain words with respect to their position inside a dependency tree, while the latter infers the emotion value of phrases and sentences in a bottom-up fashion and also takes into account symbolic hints such as interjections and emoticons. As to performance, they are on a par with each other although UPAR7 would be superior, if its recognition capabilities for Surprise would influence the performance average.

Similar to the baseline system, DM-f [32] and SWAT [18] rely exclusively on averaging word emotions as taken from their incorporated lexicons. For the SWAT system, a lexicon was trained using human-annotated news headlines. It yields reasonable performance although it is outperformed by the lingustics-based systems. The DM-f system, however, uses the (raw frequency version of the) DM lexicon as described above. Interestingly, combing this extensive lexicon with the simple average-word-emotion approach yields far better results than any other system presented so far. Thus, for this task, lexicon coverage seems to beat structural language properties to some extent.

Concerning systems using the VAD model, Calvo and Kim [10] use this dimensional model as an intermediate representation later on mapping the VAD values onto (coarse-grained) BEs (cf. Section 2.3). Therefore, they do not offer a metrical evaluation for those dimensional assessments. Leveau et al. [20], in an approach similar to ours, average Valence and Arousal values of words for French texts. Being primarily a psychological study, this work also does not offer a meaningful evaluation from an NLP point of view. In a preceding study [9], we used a less sophisticated version of our system to measure emotions in a large corpus of business reports but did not provide a metrical evaluation due to (at that time) the lack of test data. In another recent study, Preoţiuc-Pietro et al. [27] predict Valence and Arousal values in FACEBOOK posts using linear regression models with bag-of-words features. They report performance figures of $r = .65$ and .85 for Valence and Arousal, respectively.

Note that prior studies using lexicon-based methods differ in weighting procedures: some of them emphasize the emotion of a word occurring in a document using absolute term frequencies (TF) (e.g., [18]), whereas others rely on TF-IDF scores (e.g., [35]). However, no data on the impact of either one of these weighting schemes has been made available (although Staiano and Guerini [32] compared lexicons *constructed* with different weighting functions).

[4] `http://www.natcorp.ox.ac.uk/`

3 Experiments Using Dimensional Models

We start in Section 3.1 by defining a metrical criterion which guides the emotion analysis for JEMAS (Jena Emotion Analysis System),[5] our bag-of-words (BOW) engine (similar to [32] and [18]) employing the VAD model. In Section 3.2, we then evaluate JEMAS using different configurations and discuss implications of these experiments concerning metrical evaluation in Section 3.3.

3.1 Simple Metrics for Emotion Analysis

We distinguish two basic data containers. First, the set of documents (1) where λ denotes some weighting function for terms and $t_{i,j}$ denotes some morphologically normalized non-stop word term in the document-term vector for document d_i, $j = 1, ..., n$; n being the total size of the normalized vocabulary in DOC, so that $\lambda_{t_{i,j}}$ denotes the numerical weight of the *j-th* term from document d_i. Second, the VAD lexicon (2) where each emotion-sensitive lemma lex_l contained in VAD is associated with its corresponding VAD triple $\langle v_l, a_l, d_l \rangle \in \mathbb{R}^3$; each of the three components ranges in the normalized interval $[-4, 4]$, with $l = 1, .., t$; t enumerating the total size of the lexicon.

$$DOC := \{d_i = (\lambda_{t_{i,1}}, ..., \lambda_{t_{i,n}})\} \tag{1}$$

$$VAD := \{vad_l = (lex_l, \langle v_l, a_l, d_l \rangle)\} \tag{2}$$

We may then define the *Emotion Value* of each document d_i (using the projection $\pi_1(VAD) := \{lex \mid (lex, \langle v, a, d \rangle) \in VAD\}$ and the string equality function SEQ):

$$EV_{d_i} := \frac{\sum_{k=1 \,\wedge\, \exists lex_q \in \pi_1(VAD):\, SEQ(lex_q, t_{i,k})}^{n} \lambda_{t_{i,k}} \times \langle v_q, a_q, d_q \rangle}{\sum_{k=1 \,\wedge\, \exists lex_q \in \pi_1(VAD):\, SEQ(lex_q, t_{i,k})}^{n} \lambda_{t_{i,k}}} \tag{3}$$

The general purpose of the term weighting functions λ is to capture the importance a given term, $t_{i,j}$, has for a document d_i. For the following experiments, we specify two such weighting functions (although any other term weighting function for document-term vectors can be employed in this framework). The first weighting function we use, λ_1, is the absolute frequency of a term in a document, $\mathrm{TF}_{i,j}$, that is simply the count how often term $t_{i,j}$ occurs in document d_i:

$$\lambda_1 := \mathrm{TF}_{i,j} \tag{4}$$

Secondly, we use the TF-IDF metric which is the most common weighting scheme in information retrieval [21]. Let $|DOC|$ be the total number of documents in the document collection and let DF_j be the number of documents in which t_j occurs. Hence, our second weighting scheme, λ_2, is defined by the TF-IDF weight of term t_j within the entire document collection:

$$\lambda_2 := \mathrm{TF}_{i,j} \times log \frac{|DOC|}{\mathrm{DF}_j} \tag{5}$$

3.2 Evaluation of the JEMAS Emotion Analyzer

This formal sketch is flexible enough to process documents of arbitrary length, i.e. ranging from a single word to hundreds of pages of full text [9]. However, in the following experiment, we use ANET [6] as a test corpus for the JEMAS system. We transform the VAD ratings associated with the 120 short texts into the interval $[-4, 4]$, with '0' as the neutral rating point for each of the three VAD dimensions. Concerning the chosen lexicons, we decided to compare all of the three lexicons introduced in Section 2 incorporating the VAD model of emotion since they vary largely in terms of size and the underlying acquisition methodology, i.e.,

- the extended (2010-) version of ANEW [7] which—although being rather small—was compiled using a controlled experimental environment,
- the BV lexicon [2] assembled via bootstrapping from the original 1999-version of ANEW [5], and
- the WKB lexicon [39] which reproduces and extends the original ANEW by crowdsourcing.

We transform the emotion value of each lexicon entry so that they are balanced in the interval [–4,4] to simplify interpretation (in the original lexicons, they range in the interval [1,9]).

Since no data on the impact of different term-weighting schemes is available (cf. Section 2.4), we generate results for both, TF and TF-IDF schemes, for a total of six configurations of our system (one for each combination of lexicon and weighting function). Table 4 presents the evaluation results (given in Pearson's correlation) for this experiment.

Table 4. Results of the JEMAS system (Pearson's r) relative to the three VAD dimensions. Evaluation was performed against the ANET corpus with all possible combinations of lexicons and weighting functions.

	Valence		Arousal		Dominance		Avg.	
	tf	tfidf	tf	tfidf	tf	tfidf	tf	tfidf
ANEW	0.53	0.56	0.58	0.58	0.43	0.46	0.51	0.53
BV	0.67	0.68	0.49	0.48	0.66	0.65	0.61	0.61
WKB	0.70	0.71	0.63	0.64	0.59	0.59	0.64	0.65

In general, we find the correlation to the human ratings to be between $r = 0.43$ and 0.71 depending on the lexicon, the weighting function and especially the respective emotional dimension. The crowdsourced and high-volume WKB lexicon provides the best average correlation over Valence, Arousal and Dominance. The BV lexicon gets slightly worse performance figures but still mostly exceeds those that can be achieved using ANEW (except for Arousal). Hence, in terms of performance with respect to the lexicons, coverage seems to beat quality to some extent.

These findings can be further connected to the *recognition rate* of our system, i.e., the percentage of content words in a document which can be attributed an emotion value by our system, using one of the three lexicons: we obtain 42%, 95%, and 87% recognition using the ANEW, the BV, and the WKB lexicon, respectively.

The data suggest that the performance boost of BV and WKB over ANEW can be well explained by superior coverage, whereas the coverage gain BV shows in comparison with WKB seems to be more than compensated by WKB's superior quality due to human ratings as opposed to the semi-supervised approach underlying BV. Note that the BV lexicon used the 1999 version of ANEW as seed set for bootstrapping but still yields better results than the 2010 edition (which

[5] JEMAS will be publicly available on our GITHUB site `https://github.com/JULIELab`.

has more than twice the amount of entries), thus demonstrating the validity of Bestgen and Vincze's [2] bootstrapping method.

Concerning the comparison of TF versus TF-IDF weighting functions, our data (see Table 4) hint at a slight advantage when using TF-IDF scores leading to an increased correlation in seven instances while decreasing it in only two (for Arousal and Dominance using BV). Also, average performance increased by one, respectively two percent points for WKB and ANEW while it remains unchanged for BV. A possible explanation for this improvement could be that common words are emotionally rather neutral and rating consistency is rather poor for emotionally neutral words [39]. Therefore, words whose emotion values are less reliable may be attributed less relevance using TF-IDF resulting in an overall gain in performance.

Of course, our results are not directly comparable to the ones from prior evaluation rounds as shown in Table 3 due to different test corpora and models of emotion. However, it should be noted that the correlation our system obtains with human ratings for the ANET corpus (concerning the VAD emotions) widely exceeds the correlation any of the systems revealed when they are evaluated against the SEMEVAL corpus (in relation to Ekman's six basic emotions) and even exceeds human IAA for two different studies (Table 2). This result is even more exciting since our methodology resembles that of those prior systems, especially DM-f [32], which also employs a broad-coverage emotion lexicon and, in essence, averages word emotion values. We carefully interpret this observation as possibly hinting at the superiority of the VAD model (in terms of its suitability for inter-subjective and reliable assessments for humans, as well as for algorithms) compared with the BE model, a stipulation we further elaborate after the discussion of further experiments below.

Comparing our findings to those of Preoţiuc-Pietro et al. [27], it becomes apparent that performance in emotion analysis strongly depends on the specific domain, i.e., they report a performance of only $r = .113$ and .188 for Valence and Arousal, respectively, using the WKB lexicon on their FACEBOOK posts corpus (in contrast to our system performing at $r = .70$ and .65 using a very similar set-up on the ANET corpus) while linear regression models using BOW features perform at $r = .65$ and .85.

Extending the usual evaluation methodology for fine-grained emotion detection, we decided not only to measure the performance of our system with respect to Pearson's correlation but to also take into account root-mean-square error (RMSE) which is commonly used to assess the quality of a regression model. It is computed as the quadratic mean of the errors, i.e., the differences between the values predicted by the model and the values actually observed. Table 5 displays the same data as in Table 4 for RMSE instead of r.

Table 5. Results of the JEMAS system (RMSE) relative to the three VAD dimensions. Evaluation was performed against the ANET corpus with all combinations of lexicons and weighting functions.

	Valence		Arousal		Dominance		Avg.	
	tf	tfidf	tf	tfidf	tf	tfidf	tf	tfidf
ANEW	2.38	2.33	1.80	1.82	1.78	1.75	1.98	1.97
BV	2.42	2.41	2.03	2.04	1.79	1.79	2.08	2.08
WKB	2.26	2.23	2.57	2.56	1.80	1.78	2.21	2.19

The surprising result of applying RMSE for these configurations is that the relative performance of the three lexicons when compared to one another changes completely. While with r WKB outperfomed BV which itself yielded better results than ANEW, using RMSE, the order of the lexicons according to the measured performance figures is actually reversed (note that since RMSE denotes a measure of error, the lower the value the better the performance).

To further investigate this astonishing result, we plotted the data (only TF-based results) in nine scatterplots (see Figure 2) where each row (with three plots each) displays the results for one lexicon and each column depicts the results for one emotional dimension. Accordingly, a data point in a particular plot denotes the predicted value for an instance of ANET (x-axis) in one emotional dimension using one of the three lexicons and its actual value according to the human ratings (y-axis). The red lines designate the regression line (using a linear model) while the green lines (for comparison) denote a perfect agreement (predicted values equal actual values).

Building on these data visualizations, we venture to cautiously explain the opposing result in terms of r and RMSE. As can be seen, the data points scatter loosely around the regression line when using the ANEW lexicon, whereas for BV and WKB they stick considerably closer to it. Since the (vertical) distance of a data point to the regression line is related to the linear relationship between the two data series, this observation visually "explains" that r values are getting higher from the top line to the bottom line of the scatterplots.

Also, it can be seen that the slope of the regression is much steeper when using the BV and WKB lexicon. The slope of the regression line is related to the interval the predicted values are ranging in. As can be observed, x-values ranging in a small interval result in a steeper slope. This means that data points can be positioned closely to the regression line while at the same time (because of its slope) being far away from the green line (denoting a considerable difference between predicted and actual value).

For instance, in the middle column, the actual value of an instance may be, say, 3 so perfect agreement would demand for a predicted value of 3 as well (as marked by the green line). However, for such instances, our system usually predicts (approximately) a value of 0 (as can be seen) resulting in a large *squared* error. At the same time, the data point being close to the regression line contributes to a high Pearson's correlation (r). Also note that the data points for predicting Arousal with the WKB lexicon (bottom center plot) are off-center (a property most probably derived from the lexicon itself [39]) resulting in an even higher squared error.

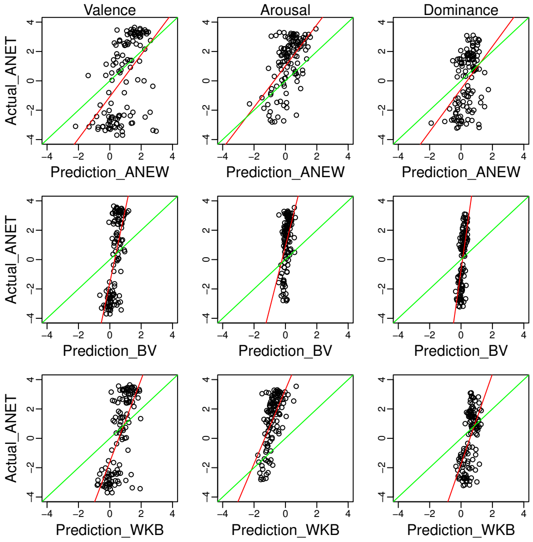

Figure 2. Scatterplotts for a graphical interpretation of the evaluation against the ANET corpus using TF weights. Each data point in each plot designates a pair of a predicted value (x-axis) and the actual value according to ANET (y-axis). The plots are grouped by lexicons used for the evaluation (row-wise) and by emotional dimensions (column-wise).

The steepness of the slopes seems to correlate with the number of entries in the lexicon used to produce the particular data, as well as with the recognition rate (see above). This seems to indicate that the bigger the lexicon, the larger the error caused by this effect could be. A possible explanation for these findings is that most of the words contained in a large emotion lexicon, in contrast to a small one, are on average less emotional (because strongly emotion-bearing lexemes will most likely already be included in a small lexicon, right from the beginning). We conclude that for high performance in terms of Pearson's correlation, the relative differences between the predicted values should be reliable but still the numeric values may differ a lot from the actual values making *any* system unreliable.

The consequences of the above interpretation may, to some extent, be dramatic. Arguably, the prevailing performance measure (Pearson's r) used up until now captures only half of our human intuition of textual emotion, i.e. how the emotion associated with one linguistic unit relates to that of another one—this aspect of a model's predictive power is captured by correlation. It does, however, not capture our ability to perceive the strength and orientation of an emotion with respect to an absolute scale (e.g., neutral arousal vs. highest arousal)—that aspect of a model's predictive power is captured by an error-based metric. While the former may be sufficient for some tasks, it may be irrelevant for others. Therefore, our findings point out that the common evaluation methodology for fine-grained emotion detection is seriously flawed which casts doubt on the validity of prior results (Table 3). Furthermore, since the phenomenon described above was observed using a lexicon-based method (it became more pronounced the larger the coverage of such a resource is), it seems quite likely that, e.g., DM-f, the best-performing system, displays a similar behavior due to the commonalities of the two approaches (ours and theirs). For future work, we therefore suggest to use RMSE as a performance measure complementary to r because taking account of error might be more relevant than the consideration of correlation for many applications and must therefore be addressed during evaluation.

3.3 A Linear Regression-Based Repair Mechanism

In a first attempt to cope with the newly discovered weaknesses of our system, we developed a simple, yet effective repair mechanism to better fit our predictions to the actual data. For each combination of lexicon, weighting function and emotional dimension according to the VAD model, we trained a linear regression model (18, in total) using the originally predicted value of the particular emotion as the only (input) feature. Training was conducted using the ANET corpus. We did not perform cross-validation because these models cannot overfit due to their simplicity. We then post-processed our data from the previous experiment using these models. Table 6 depicts the results of this experiment using RMSE as evaluation yardstick. Note that Pearson's correlation remains unchanged by this procedure.

Table 6. Evaluation result against the ANET corpus after linear regression-based repair (measurements in RMSE).

	Valence		Arousal		Dominance		Avg.	
	tf	tfidf	tf	tfidf	tf	tfidf	tf	tfidf
ANEW	2.23	2.18	1.40	1.41	1.72	1.69	1.78	1.76
BV	1.95	1.93	1.50	1.51	1.42	1.44	1.62	1.62
WKB	1.87	1.85	1.33	1.32	1.54	1.53	1.58	1.57

As can be seen, Table 6 resembles Table 5 in many key features, e.g., TF-IDF yields slightly better results than TF (demonstrating the robustness of this method). However, each single RMSE value experienced a pronounced drop of error so that the higher the error was, the more the RMSE decreased, thus rearranging the relative performance figure between the lexicons. After repair, according to RMSE measurements, WKB yields better results than BV which itself is better than ANEW. Thus, the order has been reversed in comparison to the data without repair. Furthermore, orderings are now consistent with the ones when using r as performance measure.

The visual interpretation of this method is that instead of predicting the output value of our system, we predict the point on the regression line (displayed in red in Figure 2) associated with it, that is the point above its value on the x-axis. As a result, the new regression line is identical to the line of perfect agreement (displayed in green in Figure 2). We conclude that our method yields satisfactory results despite its simplicity. Yet, the corpus we use for this experiment is quite small (120 instances) so that our method could be less effective when applied to other data sets.

4 Comparison with Categorical Systems

In previous work, we have demonstrated the practical value the VAD data our system produces may have for other areas of research, e.g., emotional portrays of enterprises based on their business and sustainability reports [9]. In contrast, the following section addresses the mapping from VAD to BE representation as a methodological exercise only for the sake of comparison since the number of directly comparable systems is otherwise extremely limited.

4.1 Mapping Emotion Models

As already discussed, being able to reliably convert between different models of emotions, such as the VAD and the BE model, yields many benefits, including better reusability of resources, as well as better means of comparing emotion detection systems using different representation schemes for emotions. Building on the work of Stevenson et al. [33], we here use their complementary BE-based emotional ratings for the ANEW lexicon to generate a variety of regression models. We start by transforming ANEW's VAD and BE ratings so that the former are balanced in the interval $[-4, 4]$, as we already did with our lexicon, and the latter span the interval $[0, 100]$ so that their interval equals that of the SEMEVAL-2007 test corpus. We used the R CARET package[6] to train linear models, SVMs with a polynomial kernel and kNN models for regression. For either way of the emotion model mapping (VAD → BE, as well as BE → VAD), we trained an independent model for each category or dimension of the emotion representation we map onto, while each category or dimension of the input representation was used as a feature.

For example, when transforming basic emotion to their VAD representation, we trained three independent models each one relying on all of the basic emotions as features. As our models are solely based on the input emotion values (not taking into account other features), they are independent from the type of stimulus eliciting the emotion, e.g. be it a word, a sentence, a text or an image. The performance of these models (obtained using 10-fold cross-validation) is summarized in Tables 7 and 8 using R^2. These values are consistent with the RMSE-based results. Note that, since each table cell represents an independent model, tuning parameter selection may differ across a particular line.

[6] http://topepo.github.io/caret/index.html

Table 7. Performance of statistical models—linear regression (lm), support vector machine with polynomial kernel (svmPoly) and k-Nearest Neighbor regression model (kNN)—for mapping VAD to BE emotion representation, measured in R^2.

	Anger	Disgust	Fear	Joy	Sadness	Avg.
lm	0.734	0.584	0.736	0.867	0.678	0.720
svmPoly	0.760	0.625	0.757	0.918	0.764	0.765
kNN	0.759	0.635	0.754	0.922	0.747	0.763

Table 8. Performance of statistical models—linear regression (lm), support vector machine with polynomial kernel (svmPoly) and k-Nearest Neighbor regression model (kNN)—for mapping BE to VAD emotion representation, measured in R^2.

	Valence	Arousal	Dominance	Avg.
lm	0.934	0.528	0.704	0.722
svmPoly	0.944	0.562	0.722	0.743
kNN	0.935	0.523	0.702	0.720

Overall, the machine learning approach gave good results with averaged R^2 ranging roughly between 72 and 77% both ways. Joy and Valence are predicted best, with values above 90%, whereas Disgust and Arousal are predicted far less accurately. Both ways, SVMs performed best. For mapping onto the BE model, kNN regression was almost equally good, whereas for mapping onto VAD emotions, surprisingly, a simple linear model outperformed kNN.

4.2 Evaluation Using Representation Mappings

In our last experiment, we use the regression models we trained for emotion representation mapping to compare the performance of the JEmAS system with prior ones in a more direct way. We use our system to predict VAD ratings for the SemEval test corpus (supplied only with BE annotations) employing the WKB lexicon and the TF-IDF weighting scheme, since this configuration obtained the best performance. The newly developed repair mechanism was not included, since the performance figures of the other systems are reported only using r values on which this method has no effect. The resulting VAD predictions were mapped onto basic emotions using the SVMs we trained on the Anew lexicon. Finally, we computed Pearson's correlation between the resulting BE values and the human ratings provided for the SemEval corpus. The results of this set-up are depicted in Table 9.

Table 9. Results of evaluating the JEmAS system against the SemEval-2007 corpus after mapping its VAD output onto basic emotions. Improvements over the formerly best systems (per emotion category, cf. Table 3) in bold face.

Anger	Disgust	Fear	Joy	Sadness	Surprise	**Avg.**
.399	**.252**	.440	**.469**	.366	—	.419

With a mean performance of $r = .419$ (considering Anger, Fear, Joy and Sadness—these are the categories each system covers) the JEmAS system yields state-of-the art performance for three out of six emotion categories (namely Anger, Disgust and Joy) overall clearly out-performing any existing system but one (DM-f) even *after* applying the imperfect transformation into BE representations. Its relatively high performance seems in some categories (e.g., Disgust) highly counter-intuitive taking into account that our system has no direct or apparent way of measuring these categories while all the other systems have mechanisms (e.g., keywords) specifically supplied for addressing them. Obviously the favorable evaluation results our system achieves in terms of VAD (Table 4) were not mainly an effect due to corpus bias but arguably, since it is still among the top-performers after emotion representation mapping, it must be considered on a par with, if not superior, to the best-performing present system. Note that the results would be even more favorable for JEmAS, if performance were reported in an error-based metric due to our repair mechanism for the large-lexicon bias (cf. Section 3.3).

5 Conclusions

In this work, we addressed multiple central issues of fine-grained emotion analysis—the task of predicting the associated emotion given a linguistic unit such as a sentence or a text. A fine-grained analysis differs from its coarse-grained counterpart by translating into a regression, rather than a classification problem. We offered a critical comparison of the two prevailing models of emotion in computational approaches—Russell and Mehrabian's Valence-Arousal-Dominance model and Ekman's Basic Emotion model—pointing out problematic aspects of the latter, especially in a regression set-up.

Building on these theoretical considerations, we here presented JEmAS, the first evaluated system measuring VAD-based emotions. As this system uses a lexicon-based approach, evaluation was carried out incorporating three different lexicons and two different term weighting function for a total of six configurations. Despite the simplicity of our approach, it yields satisfying performance figures of up until $r = .65$ (average over Valence, Arousal, and Dominance). Instead of solely using Pearson's correlation as performance metric, the common basis for evaluation, we, additionally, introduced RMSE to evaluate emotion regression. The surprising result of comparing both metrics was that under both criteria performance orderings of the configurations were basically reversed depending on the lexicon being used.

A graphical analysis hinted at a reasonable explanation that, while association (measured in r) of predicted and actual values typically increases with lexicon coverage (assuming constant lexicon quality), the quadratic mean of the errors (RMSE) increases as well. As a consequence, our data indicate that using a high coverage lexicon may result in emotion predictions being fairly reliable *relative to one another*, but unreliable *relative to the orientation and absolute value* of the actual data. Since prior systems are most probably also affected by this bias, our findings indicate a severe problem for the commonly shared evaluation methodology. In a first attempt to compensate for this effect, we trained simple linear regression models to better fit our predictions to actual data resulting in a strong decrease of errors.

Since there are no directly comparable systems to JEmAS, the second half of our experiments addressed means of relating our findings more closely to prior BE-based systems. We did that by introducing a novel method of mapping between both emotion representations. That allowed us to compute VAD-values for the prevailing BE test corpus and to, then, translate our VAD output to BE representation and compare it to human judgment. Even after this imperfect (and therefore performance-reducing) mapping, our system still outperformed any prior system in three out of six emotion categories, over-all scoring on second rank (measured in r). However, existing systems do not compensate for the large-lexicon bias suggesting that our system, and its underlying methodological design decisions, may probably be superior, in terms of RMSE, at least.

ACKNOWLEDGEMENTS

We would like to thank The Center for the Study of Emotion and Attention, University of Florida for granting us access to their resources Anew and Anet on which a large proportion of this work relies.

REFERENCES

[1] Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley, 'The expression of emotions in 20th century books', *PLoS ONE*, **8**(3), e59030, (2013).

[2] Yves Bestgen and Nadja Vincze, 'Checking and bootstrapping lexical norms by means of word similarity indexes', *Behavior Research Methods*, **44**(4), 998–1006, (2012).

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng, 'TWITTER mood predicts the stock market', *Journal of Computational Science*, **2**(1), 1–8, (2011).

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

[5] Margaret M. Bradley and Peter J. Lang, '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, (1999).

[6] Margaret M. Bradley and Peter J. Lang, 'Affective norms for English text (ANET): Affective ratings of text and instruction manual', Technical Report D-1, University of Florida, Gainesville, FL, (2007).

[7] Margaret M. Bradley and Peter J. Lang, 'Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings', Technical Report C-2, University of Florida, Gainesville, FL, (2010).

[8] Joost Broekens, 'In defense of dominance: PAD usage in computational representations of affect', *International Journal of Synthetic Emotions*, **3**(1), 33–42, (2012).

[9] Sven Buechel, Udo Hahn, Jan Goldenstein, Sebastian G. M. Händschke, and Peter Walgenbach, 'Do enterprises have emotions?', in *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*, pp. 147–153, (2016).

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

[11] François-Régis Chaumartin, 'UPAR7: A knowledge-based system for headline sentiment tagging', in SEMEVAL-2007 — *Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 422–425, (2007).

[12] Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman, 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, **41**(6), 391–407, (1990).

[13] Bart Desmet and Véronique Hoste, 'Emotion detection in suicide notes', *Expert Systems with Applications*, **40**(16), 6351–6358, (2013).

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

[15] Petr Hájek, Vladimír Olej, and Renáta Myšková, 'Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making', *Technological and Economic Development of Economy*, **20**(4), 721–738, (2014).

[16] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu, 'EMOTEX: Detecting emotions in TWITTER messages', in *Proceedings of the 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference. Stanford University, CA, USA, May 27-31, 2014*, pp. 27–31, (2014).

[17] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht-Ecklundt, Harald C. Traue, and Henrik Kessler, 'Mapping discrete emotions into the dimensional space: An empirical approach', in *SMC 2012 — Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics. Seoul, Korea, 14-17 October 2012*, pp. 3316–3320, (2012).

[18] Phil Katz, Matthew Singleton, and Richard Wicentowski, 'SWAT-MP: The SEMEVAL-2007 systems for Task 5 and Task 14', in SEMEVAL-2007 — *Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 308–313, (2007).

[19] Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, and Andrés Montoyo, 'UA-ZBSA: A headline emotion classification through Web information', in SEMEVAL-2007 — *Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 334–337, (2007).

[20] Nicolas Leveau, Sandra Jhean-Larose, Guy Denhière, and Ba-Linh Nguyen, 'Validating an interlingual metanorm for emotional analysis of texts', *Behavior Research Methods*, **44**(4), 1007–1014, (2012).

[21] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

[22] Myriam D. Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen, 'Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text', *IEEE Transactions on Affective Computing*, **5**(2), 101–111, (2014).

[23] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, 'Affect Analysis Model: Novel rule-based approach to affect sensing from text', *Natural Language Engineering*, **17**(1), 95–135, (2011).

[24] Andrew Ortony, Gerald L. Clore, and Allan M. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

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

[26] R. W. Picard, *Affective Computing*, MIT Press, Cambridge/MA, 1997.

[27] Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman, 'Modelling valence and arousal in FACEBOOK posts', in *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis @ NAACL-HLT 2016. San Diego, CA, USA, June 16, 2016*, pp. 9–15, (2016).

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

[29] Klaus R. Scherer, 'Psychological models of emotion', in *The Neuropsychology of Emotion*, ed., Joan C. Borod, 137–162, Oxford University Press, Oxford, U.K.; New York, N.Y., (2000).

[30] Hashim Sharif, Fareed Zaffar, Ahmed Abbasi, and David Zimbra, 'Detecting adverse drug reactions using a sentiment classification framework', in *Proceedings of the 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University, CA, USA, May 27-31, 2014*, (2014).

[31] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng, 'Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks', in *EMNLP 2008 — Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, USA, October 25-27, 2008*, pp. 254–263, (2008).

[32] Jacopo Staiano and Marco Guerini, 'DEPECHE MOOD: A lexicon for emotion analysis from crowd annotated news', in *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, June 23-25 2014*, volume 2: Short Papers, pp. 427–433, (2014).

[33] Ryan A. Stevenson, Joseph A. Mikels, and Thomas W. James, 'Characterization of the affective norms for English words by discrete emotional categories', *Behavior Research Methods*, **39**(4), 1020–1024, (2007).

[34] Carlo Strapparava and Rada Mihalcea, 'SEMEVAL-2007 Task 14: Affective text', in SEMEVAL-2007 — *Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, June 23-24, 2007*, pp. 70–74, (2007).

[35] Carlo Strapparava and Rada Mihalcea, 'Learning to identify emotions in text', in *SAC 2008 — Proceedings of the 2008 ACM Symposium on Applied Computing. Fortaleza, Ceará, Brazil, March 16-20, 2008*, pp. 1556–1560, (2008).

[36] Carlo Strapparava and Alessandro Valitutti, 'WORDNET-AFFECT: An affective extension of WORDNET', in *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation. In Memory of Antonio Zampolli. Lisbon, Portugal, 24-30 May, 2004*, volume 4, pp. 1083–1086, (2004).

[37] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock, 'The affective weight of lexicon', in *LREC 2006 — Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 22-28 May, 2006*, pp. 423–426, (2006).

[38] Tony Veale and Yanfen Hao, 'Detecting ironic intent in creative comparisons', in *ECAI 2010 — Proceedings of the 19th European Conference on Artificial Intelligence. Lisbon, Portugal, 16-20 August 2010*, eds., Helder Coelho, Rudi Studer, and Michael J. Wooldridge, number 215 in Frontiers in Artificial Intelligence and Applications, pp. 765–770, Amsterdam, The Netherlands, (2010). IOS Press.

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