Representation Mapping:
A Novel Approach to Generate High-Quality Multi-Lingual Emotion Lexicons

Sven Buechel & Udo Hahn

{sven.buechel|udo.hahn}@uni-jena.de
Jena University Language and Information Engineering (JULIE) Lab
Fürstengraben 27, 07743 Jena, Germany
http://www.julielab.de

Abstract

In the past years, sentiment analysis has increasingly shifted attention to representational frameworks more expressive than semantic polarity (being positive, negative or neutral). However, these richer formats (like Basic Emotions or Valence-Arousal-Dominance, and variants therefrom), rooted in psychological research, tend to proliferate the number of representation schemes for emotion encoding. Thus, a large amount of representationally incompatible emotion lexicons has been developed by various research groups adopting one or the other emotion representation format. As a consequence, the reusability of these resources decreases as does the comparability of systems using them. In this paper, we propose to solve this dilemma by methods and tools which map different representation formats onto each other for the sake of mutual compatibility and interoperability of language resources. We present the first large-scale investigation of such representation mappings for four typologically diverse languages and find evidence that our approach produces (near-)gold quality emotion lexicons, even in crosslingual settings. Finally, we use our models to create new lexicons for eight typologically diverse languages.

1. Introduction

In the past two decades, the NLP-based analysis and prediction of affective states, as performed by sentiment analysis systems, has received enormous interest (Liu, 2015). Starting with simple positive-negative polarity distinctions on the word or text level (Hatzivassiloglou and McKeown, 1997; Pang et al., 2002), research in sentiment analysis has since then shifted towards more nuanced and challenging tasks, e.g., sentiment compositionality (Socher et al., 2013), aspect-level assessments (Schouten and Frasincar, 2016) or stance detection (Sobhani et al., 2016). In parallel, psychologically more advanced and more expressive *representation formats* for affective states have been proposed, like Basic Emotions (Ekman, 1992) or Valence-Arousal-Dominance (Bradley and Lang, 1994). However, there is currently no consensus in the literature what scheme should be used as a common ground. Rather, there are a multitude of competing formats often motivated by the needs of concrete applications or the availability of user-labeled social media data (Desmet and Hoste, 2013; Li et al., 2016).

While such decisions for a specific format may be perfectly reasonable in a specific research setting, on the flip-side, this proliferation of competing formats may seriously hamper progress in sentiment analysis for two reasons, at least. First, language resources are less reusable (if at all) as gold standards and, second, with the growing number of representation formats meaningful comparisons between predictive systems become harder (if not impossible).

One way to resolve this dilemma is to develop techniques to automatically translate between such formats. This task of *emotion representation mapping* (EMOMAP) was introduced only very recently to NLP by Buechel and Hahn (2017b). Their work came up with an emotion-labeled corpus which, in part, is annotated with two different emotion formats both being highly predictive for each other. In a follow-up study, Buechel and Hahn (2017a) examined

the potential of EMOMAP as a substitute for manual annotation, yet their comparison was restricted to only two emotion lexicons. Comparable work has (to the best of our knowledge) only been done in psychology. However, this stream of work does not target the goal of predictive modeling (Stevenson et al., 2007; Pinheiro et al., 2017). In NLP, a task related to EMOMAP is *emotion prediction* on the level of words, sentences, or texts (Wang et al., 2016; Sedoc et al., 2017) where, in contrast to EMOMAP, the target unit does not already need to bear annotations from another format. Thus, emotion prediction algorithms constitute a reasonable baseline for EMOMAP (see Section 4.).

This contribution puts emphasis on emotion lexicons developed in psychology. Although highly relevant for sentiment analysis, those resources have mostly been neglected by NLP researchers as the discussion of related work in Section 2. reveals. Making use of this valuable work, we here conduct the first thorough evaluation of EMOMAP for emotion lexicon construction on four typologically diverse languages and find strong evidence that the quality of the output we generate is on a par with a gold standard when compared to human performance (see Section 4.). Finally, we exploit our models to create novel emotion lexicons for eight different languages (including low-resource ones; Section 5.). The lexicons as well as the source code for building them are publicly available (see Section 6.).

2. Data

Models of emotion are typically subdivided into *discrete* (or *categorical*) and *dimensional* ones (Stevenson et al., 2007; Calvo and Mac Kim, 2013). Discrete models are centered around particular sets of emotional categories deemed fundamental. Ekman (1992), for instance, identifies six *Basic Emotions* (Joy, Anger, Sadness, Fear, Disgust and Surprise). In contrast, dimensional models consider emotions to be composed out of several influencing factors (mainly
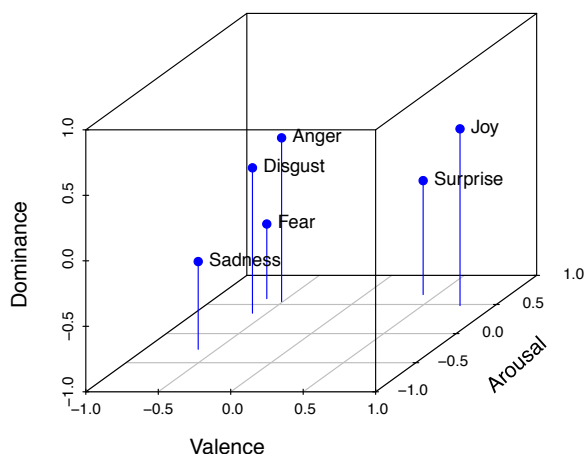
Figure 1: Affective space spanned by the Valence-Arousal-Dominance model, together with the position of six Basic Emotions; positions determined by Russell and Mehrabian (1977); figure adapted from Buechel and Hahn (2016).

Reference	Lang.	Format	# Entries
Warriner et al. (2013)	en	VAD	13,915
Stevenson et al. (2007)	en	BE5	1,034
Bradley and Lang (1999)	en	VAD	1,034
Stadthagen-Gonzalez et al. (2017)	es	VA	14,031
Ferré et al. (2017)	es	BE5	2,266
Guasch et al. (2015)	es	VA	1,400
Redondo et al. (2007)	es	VAD	1,034
Hinojosa et al. (2016a)	es	VA+BE5	875
Hinojosa et al. (2016b)	es	+D	875
Võ et al. (2009)	de	VA	2,902
Briesemeister et al. (2011)	de	BE5	1,958
Schmidtke et al. (2014)	de	VAD	1,003
Kanske and Kotz (2010)	de	VA	1,000
Imbir (2016)	pl	VAD	4,905
Riegel et al. (2015)	pl	VA	2,902
Wierzba et al. (2015)	pl	BE5	2,902
Yu et al. (2016)	zh	VA	2,802
Yao et al. (2017)	zh	VA	1,100
Monnier and Syssau (2014)	fr	VA	1,031
Ric et al. (2013)	fr	V+BE5	524
Moors et al. (2013)	nl	VAD	4,299
Sianipar et al. (2016)	id	VAD	1,490
Palogiannidi et al. (2016)	gr	VAD	1,034
Montefinese et al. (2014)	it	VAD	1,121
Soares et al. (2012)	pt	VAD	1,034
Eilola and Havelka (2010)	fi	VA	210
Davidson and Innes-Ker (2014)	sv	VA	100

Table 1: List of VA(D) and BE5 lexicons with empirically gathered ratings from human subjects; including reference, language code (according to ISO 639-1), emotion representation format, and number of lexical entries.

Word	V	A	D	J	A	S	F	D
sunshine	8.1	5.3	5.4	4.2	1.2	1.3	1.3	1.2
terrorism	1.6	7.4	2.7	1.1	3.1	3.4	3.7	2.7
orgasm	8.0	7.2	5.8	4.2	1.3	1.3	1.5	1.2

Table 2: Three lexical items and their emotion values in VAD (second column group) and BE5 (third column group) format. VAD scores are taken from Warriner et al. (2013), BE5 scores were automatically derived (see Section 5.).

two or three). These are often referred to as *Valence* (corresponding to the concept of polarity), *Arousal* (a calm–excited scale), and *Dominance* (perceived degree of control over a (social) situation)—the VAD model (see Figure 1 for an illustration of the relationship between VAD dimensions and Basic Emotion categories). The last dimension, Dominance, is sometimes omitted, leading to the VA model.

In contrast to NLP where many different formats are being used lexical resources in psychology almost exclusively subscribe to VA(D) or Basic Emotions (typically omitting Surprise; the BE5 format). Over the years, a considerable number of resources built on these premises have emerged from psychological research labs for various languages. Table 1 enumerates published resources based on these two approaches (27 in total covering 13 languages, including low-resource ones such as Finnish and Indonesian). To the best of our knowledge, the vast majority of them has neither been used nor referenced in NLP research.

In this paper, we restrict ourselves to the VAD and BE5 format. In more detail (following the conventions of our emotion lexicons), each VAD dimension receives a value from the interval $[1, 9]$ where '1' means "most negative/calm/submissive", '9' means "most positive/excited/dominant" and '5' means "neutral". Conversely, values for BE5 categories range in the interval $[1, 5]$ where '1' means "absence" and '5' means "most extreme" expression of the respective emotion.[1] Consequently, the VAD and BE5 formats are conceptually different from one another insofar as VAD dimensions are bi-polar, whereas BE5 categories are uni-polar.

Our work is based on the condition that some pairs of data sets in Table 1 are complementary in the sense that, when combining these lexicons, a subset of the entries they contain are then described according to *both* emotion formats, VAD *and* BE5. This condition is illustrated for three lexical items in Table 2.

From the resources listed in Table 1, we identified such complementary pairs and merged them into four lexicons which serve as gold data for the subsequent experiments:

- **English**: Bradley and Lang (1999) intersected with Stevenson et al. (2007) yielded 1,034 overlapping entries.
- **Spanish**: Redondo et al. (2007) intersected with Ferré et al. (2017) yielded 1,012 overlapping entries.
- **Polish**: Imbir (2016) intersected with Wierzba et al. (2015) yielded 1,272 overlapping entries.
- **German**: Schmidtke et al. (2014) intersected with Briesemeister et al. (2011) yielded 318 overlapping entries.

---

[1] Although these intervals are fairly well established conventions, in some data sets different rating scales are used, nevertheless. In these cases, we linearly transformed the ratings so that they match the defined intervals.
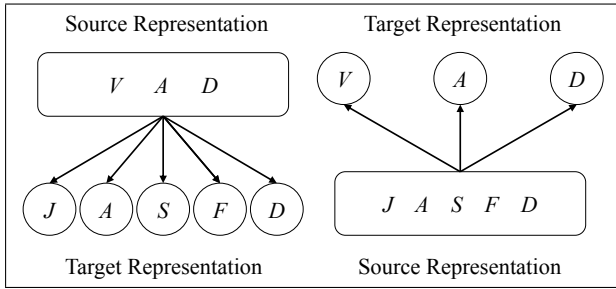
Figure 2: Illustration of the two representation mapping procedures: VAD2BE5 (left) vs. BE52VAD (right). Each arrow represents an individual kNN model.

3. Method

Given an emotion lexicon in VAD format, our goal is to map its ratings onto the BE5 format and vice versa. We employ a simple, yet surprisingly efficient, method proposed by Buechel and Hahn (2017b): For each of the dimensions or categories of the target representation (VAD or BE5, respectively), we train a single supervised model which employs each of the dimensions/categories of the source representation as features (e.g., *one* model to predict Joy, given Valence, Arousal, and Dominance scores as input; see Figure 2 for a graphical illustration of the general scheme).

In a pilot study, we compared different learning algorithms including linear regression, *k* nearest neighbor regression (kNN), support vector regression (using different kernels), random forests, as well as feed-forward neural networks. To our surprise, all of them performed equally well (with only negligible differences). Thus, kNN was selected due to its simplicity.[2] Note that the feature set is extremely small (either three or five variables for mapping *onto* BE5 or VAD, respectively) so that using more complex methods (e.g., more sophisticated neural architectures) seems a waste of efforts.

4. Experiments

We here present the first large-scale evaluation of emotion representation mapping (EMOMAP). Our methodology, at the same time, leads to the automatic construction of emotion lexicons for four typologically diverse languages. We consider one monolingual and two crosslingual set-ups, i.e., training and testing data from the same or different language(s), respectively. Those three different mapping strategies are illustrated in Figure 3.

The performance of the EMOMAP approach will be measured as Pearson correlation (r) between our automatically predicted values and human gold ratings. In general, the Pearson correlation between two data series $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_n$ takes values between $+1$ (perfect positive correlation) and -1 (perfect negative correlation). It is computed as

$$r_{xy} := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x}$ and $\bar{y}$ denote the mean values for X and Y, respectively.

---

[2]We use the `scikit-learn.org` implementation.

These measurements will then be compared, first, with the current state-of-the-art in word-level emotion prediction (as baseline), and, second, with human inter-study reliability (as ceiling). Both comparisons will, for different reasons, be limited to the VAD model.

4.1. Baseline and Ceiling

Word-level emotion prediction (automatically deriving the emotion of a word from scratch; see Section 1.) serves as a reasonable baseline since it produces the same output as EMOMAP, yet does not require the target words to have already been annotated in a different emotion format (other than the output representation).

Sedoc et al. (2017) evaluated their approach to word-level emotion prediction on the data set compiled by Bradley and Lang (1999) using 10-fold cross-validation. They report measurements of $r = 0.806$ for Valence and $r = 0.615$ for Arousal. Concerning the other affective dimensions and categories, we are not aware of any other system predicting numerical scores for them. Thus, we will restrict our comparison to Valence and Arousal.

For comparison against the human ceiling, we found eight pairs of emotion lexicons with partially overlapping entries distributed over four languages. For each of these pairs, we computed their inter-study reliability (ISR), i.e., the Pearson correlation between the ratings from the two respective studies for each affective dimension (see Table 3). Again, because we only found lexicons with overlapping VAD (not BE5) entries, we restrict this comparison to VAD representations.

We stipulate that for all ISR values from Table 3, the *minimum* for each affective dimension constitutes the most relevant score of comparison. The rationale for this assumption is as follows: If our approach happens to outperform this minimal value, one cannot be certain that manual annotation leads to better results than using our automatic procedure. In this situation, we assume that the computational approach would almost always be preferred over manual annotation efforts. Accordingly, the following correlation values were identified as minimum inter-study reliabilities: $r = .948$ for Valence, $r = .709$ for Arousal, and $r = .794$ for Dominance (henceforth, jointly referred to as ISR$_{min}$).

Note that comparing against the ISR is a much harder test than comparing against inter-annotator agreement (IAA): Since the former is based on the mean rating of many raters, this aggregated judgment is more stable than individual ratings, thus resulting in higher correlation values compared to its IAA counterpart.[3]

4.2. Monolingual Evaluation

The first part of our analysis concerns train and test data originating from the same language (respectively data set); see left part of Figure 3. For each of our (language-wise) four gold lexicons (cf. the final paragraph of Section 2.), we train kNN models to map between VAD and BE5 representations back and forth according to the scheme from Fig-

---

[3]For numerical emotion ratings, IAA is typically computed in a leave-one-out fashion and can thus be interpreted as how well a single human annotator predicts the gold value (Strapparava and Mihalcea, 2007; Buechel and Hahn, 2017b).

	Val	Aro	Dom	#Overlap
Imbir (2016) vs. Riegel et al. (2015)	**.948**	.733	—	1,272
Guasch et al. (2015) vs. Stadthagen-Gonzalez et al. (2017)	.949	**.875**	—	1,298
Bradley and Lang (1999) vs. Warriner et al. (2013)	.952	.760	**.794**	1,027
Guasch et al. (2015) vs. Hinojosa et al. (2016a)	.968	.777	—	134
Guasch et al. (2015) vs. Redondo et al. (2007)	.969	.844	—	316
Hinojosa et al. (2016a) vs. Stadthagen-Gonzalez et al. (2017)	.970	**.709**	—	636
Schmidtke et al. (2014) vs. Kanske and Kotz (2010)	.971	.788	—	169
Redondo et al. (2007) vs. Stadthagen-Gonzalez et al. (2017)	**.976**	.755	—	1,010

Table 3: Inter-study reliabilities between different data sets (measured in r). Minimum and maximum values for each VAD dimension (respectively) in bold.
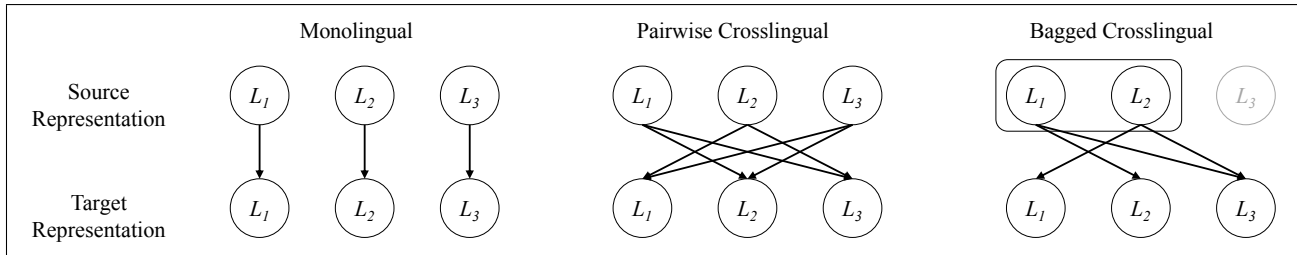

Figure 3: Illustration of the three mapping strategies applied in Section 4. exemplified for the languages L_1, L_2 and L_3.

ure 2. Training and testing was done using 10-fold cross-validation (9:1 train/test split). The k-parameter was fixed to 20, based on a pilot study (eliminating the need for a dev set). The results are presented in Table 4, upper section.

As can be seen, the outcome is overall favorable for our approach. In general, it works about equally well in both mapping directions (VAD2BE5 and BE52VAD) with average values (over VAD dimensions and BE5 categories, respectively) of $r \geq 77\%$. The results on the English and Spanish gold lexicons are better than for the Polish, yet worst for the German one (which is also the smallest). In comparison with the baseline (see above; English data set only), our EMOMAP approach performs more than 15 percentage points better for Valence and more than 10 percentage points better for Arousal. Even more surprisingly, compared to the human ceiling, we find that our approach outperforms the ISR$_{min}$ in 9 out of 12 cases (again, only failing to do so on the Polish and German data set). In those 9 cases where we outperformed the ISR$_{min}$, we conducted a one-tailed one sample t-test based on the 10 individual cross-validation results (Dietterich, 1998) finding significant differences in 6 of these cases ($p < .05$; marked with asterisk in Table 4).

We conclude that, in the monolingual set-up, EMOMAP performs on a par with (if not superior to) manual annotation for mapping *onto* VAD. Thus, its results can be considered as true gold data. For BE5, we cannot draw the same conclusion due to a lack of data on inter-study reliability. However, since the performance figures for the VAD2BE5 mapping are equally high, we may quite safely assume that the Basic Emotion ratings can be attributed high quality as well.

4.3. Pairwise Crosslingual Evaluation

In the two crosslingual set-ups (training and test data drawn from different languages, respectively data sets), we make use of the fact that our models do not rely on any language-specific information since the categories/dimensions describe (supposedly universal) affective states rather than linguistic entities. Thus, models trained on one language could, in theory, be applied to another without any adaptation.

Let us, first, address *pairwise* comparisons. That is, for each language, we train our kNN models on the entirety of the respective data set and then test on all the remaining languages individually (illustrated in Figure 3; resulting in a total of 12 language pairs). Since, this set-up uses fixed training and test sets, there is no need for cross-validation. The results are given in Table 4 (middle section).

Overall, the values remain astonishingly high. As can be seen, for mapping BE52VAD, the results are quite favorable for Valence with correlation values ranging well above 90% of correlation. On this dimension, our approach still outperforms the baseline by over a 15%-points margin and even surpasses the human ceiling in more than half of the cases, five of them being statistically significant. Since different from Section 4.2., we now have a fixed test set, we use a one-tailed z-test ($p < .05$) based on z-transformed correlation values (Cohen, 1995).

In contrast to Valence, the performance for Arousal and Dominance may suffer quite substantially in the crosslingual approach, depending on the combination of training and testing languages. While there is almost no performance loss for combinations of English and Spanish, the correlation decreases the most for combinations of Polish and German (especially for predicting Dominance), possibly due to data sparsity of the lexicons involved.

This outcome led us to conclude that the relationship between VAD and BE5 ratings is not fully constant across different languages (respectively data sets). Rather it seems to depend on subtle semantic differences between the translational equivalents of the affective dimensions/categories,

Experiment	Language	Val	Aro	Dom	Av$_{VAD}$	Joy	Anger	Sadn	Fear	Disg	Av$_{BE5}$
monolingual	English	.966*	.723	.833*	**.841**	.958	.870	.864	.864	.790	**.869**
	Spanish	.970*	.736	.855*	**.854**	.957	.847	.828	.870	.744	**.849**
	Polish	.944	.761*	.740	**.815**	.932	.845	.803	.784	.814	**.836**
	German	.950	.762*	.637	**.783**	.923	.793	.680	.851	.602	**.770**
crosslingual (pairwise)	es2en	.963*	.714	.794	**.824**	.948	.830	.853	.835	.780	**.849**
	pl2en	.962*	.598	.776	**.778**	.955	.845	.836	.832	.765	**.847**
	de2en	.952	.445	.762	**.720**	.952	.861	.836	.855	.746	**.850**
	en2es	.966*	.737*	.811	**.838**	.948	.791	.806	.826	.694	**.813**
	pl2es	.961*	.634	.701	**.765**	.941	.744	.763	.766	.665	**.776**
	de2es	.959*	.498	.842*	**.767**	.942	.794	.785	.839	.640	**.800**
	en2pl	.938	.655	.653	**.749**	.924	.816	.800	.751	.795	**.817**
	es2pl	.934	.663	.552	**.717**	.918	.755	.762	.653	.768	**.771**
	de2pl	.920	.674	.497	**.697**	.914	.815	.759	.700	.739	**.785**
	en2de	.940	.615	.583	**.713**	.915	.789	.678	.849	.584	**.763**
	es2de	.953	.618	.645	**.739**	.904	.789	.692	.840	.579	**.761**
	pl2de	.934	.691	.358	**.661**	.907	.768	.655	.788	.529	**.730**
crosslingual (bagged)	English	.963*	.714	.794	**.824**	.948	.830	.853	.835	.780	**.849**
	Spanish	.966*	.737*	.811	**.838**	.948	.791	.806	.826	.694	**.813**
	Polish	.939	.645	.629	**.738**	.926	.781	.780	.700	.769	**.791**
	German	.949	.635	.632	**.739**	.917	.799	.692	.844	.551	**.761**

Table 4: Results of the monolingual (Section 4.2.) and crosslingual (Sections 4.3. and 4.4.) evaluation in Pearson's r. Language 'a2b' denotes mapping from language *a* (source) to language *b* (target). Significant values are marked with '*' (compared to ISR$_{min}$; $p < .05$; VAD only), averages over VAD and BE5 (respectively) in bold.

cultural differences, or variations in the annotation guidelines, suggesting that the above assumption of language independence (not so surprisingly) may not fully hold.

In contrast to these partly inconclusive results, the outcome for mapping VAD2BE5 is much more favorable for EMOMAP and easy to describe. Compared to the monolingual set-up (relative to the target language), the drop of the average performance amounts to only a few percentage points (< 5 in most cases). Thus, the predictions for BE5 are much more robust compared to the VAD predictions which might be an effect of the respective source representation.

We conclude that in the pairwise crosslingual set-up, EMOMAP still performs really well in many cases. Yet depending on the language pair, the performance may degrade (much more severely so for mapping BE52VAD).

4.4. Bagged Crosslingual Evaluation

As evident from the last section, the performance of our mapping approach may vary depending on source and target language. However, different from the last experiment, when constructing new emotion lexicons in a crosslingual fashion, there is no need to restrict the training set to only one language. Instead, because no language-specific features are used, we may merge training data from multiple languages if this leads to a more robust predictive model. However, since not all languages (respectively data sets) seem to match well, there is no guarantee that more data sets always help boosting performance. In line with these considerations, the goal of the last experiment is to identify the best group of training data sets for automatically creating novel lexicons and to estimate their quality.

For each combination of gold lexicons of bag sizes two[4] to four and each target language, we train our models on the entirety of the bag of training data (except the one designated for testing, should it also belong to the training data) and then test on the target language; see the third data scenario in the right part of Figure 3.

In line with Section 4.3., we found that the average performance over VAD behaved less robust across different combinations of training data (ranging between $r = .730$ to .786) compared to BE5 ($r = .771$ to .806). Concerning the average over *all* emotions, the combination of Polish and German, once again, performed worst ($r = .755$), whereas the combination of English and Spanish worked best ($r = .794$; i.e., for testing on English, Spanish was used for training and vice versa, while for testing on the remaining languages, training was done on English *and* Spanish). Consequently, this combination of gold data was used for creating novel lexicons in the crosslingual set-up (see Section 5.).

Table 4 (bottom section) displays the results of this crosslingual experiment for the best performing bag of training data (comprising the English and Spanish gold lexicons, only). Thus, these performance data serve as an estimate of the quality of the novel emotion lexicons presented in Section 5. Overall, we find that the results are again favorable for our approach. The correlation with the English and Spanish data set, not surprisingly, is stronger than with the Polish and German one, confirming that these data sets form a better basis for generalization (the effect being more obvious for VAD than for BE5). In general, Valence, Joy, and

---

[4]If only one lexicon would be used for training, this one could not also be used for testing, thus making the different combinations incomparable.

	Language	#VAD	#BE5
monolingual	English		12,888
	Spanish	1,254	
	German	1,641	683
	Polish		3,633
crosslingual	Italian		1,121
	Portuguese		1,034
	Dutch		4,299
	Indonesian		1,490

Table 5: Automatically constructed gold quality lexicon resources, '#' indicates the number of previously unrated lexical units for a specific representation format.

	Joy	Anger	Sadness	Fear	Disgust
Mean	2.11	1.62	1.61	1.66	1.59
Median	1.86	1.38	1.38	1.42	1.37
Min	1.07	1.14	1.21	1.17	1.11
Max	4.40	3.38	3.81	3.74	3.26
StDev	0.79	0.47	0.47	0.49	0.47

Table 6: Descriptive statistics for the automatically constructed English BE5 lexicon.

Anger can be predicted with consistently high correlation, whereas for the other dimensions/categories we find occasional negative outliers.

Comparing our results to human reliability (in VAD only), we find that our models are superior to human ISR_{min} in 7 from 12 cases (including all cases on the English and Spanish data set). In 3 of these cases, the difference is statistically significant ($p < .05$). In comparison to the baseline (on English, VA only), our approach still clearly outperforms state-of-the-art word-level emotion prediction by a 15 and 10 percentage point margin for Valence and Arousal, respectively.

We conclude that even if no gold data for a given language are available, EMOMAP still performs comparably to human reliability when utilizing appropriate sets of training data. Some dimensions and categories seem to be reliable across data sets, whereas for others the performance may degrade, depending on the target data set. Yet, the lexicons derived in this set-up can still be attested *near-gold* quality.

Joy	Anger	Sadness	Fear	Disgust
christmas	*killer*	*chemo*	*insanity*	*felony*
happiness	*gang*	*worthless*	*motherfucker*	*enraged*
magical	*revenge*	*gonorrhea*	*terrorism*	*traitorous*
fun	*die*	*nausea*	*attacker*	*dishonesty*
enjoyment	*massacre*	*virus*	*bullshit*	*chauvinist*
bonus	*attacker*	*amputation*	*murderous*	*mistrust*
oasis	*sue*	*unhappiness*	*dangerous*	*gory*
fantastic	*hijacker*	*unsanitary*	*tragedy*	*hostile*
happy	*nigger*	*molester*	*arrest*	*racist*
sunshine	*penniless*	*lynching*	*rape*	*cellulite*

Table 7: Top 10 entries per Basic Emotion in automatically constructed English BE5 lexicon.

	V	A	D	J	A	S	F	D
V	-	−.18	+.72	+.92	−.83	−.82	−.75	−.87
A	-	-	−.18	−.03	+.58	+.46	+.67	+.41
D	-	-	-	+.68	−.66	−.76	−.68	−.61
J	-	-	-	-	−.66	−.61	−.59	−.69
A	-	-	-	-	-	+.92	+.95	+.91
S	-	-	-	-	-	-	+.91	+.85
F	-	-	-	-	-	-	-	+.82
D	-	-	-	-	-	-	-	-

Table 8: Correlation matrix (in r) for automatically constructed English BE5 (JASFD) lexicon combined with the data by Warriner et al. (2013) (VAD).

5. Construction of New Emotion Lexicons

After the positive evaluation of EMOMAP for four typologically diverse languages, our main contribution is to apply the created models to a wide variety of data sets which so far bear emotion ratings for *one* format only (either VAD or BE5). Based on our preceding experiments, we claim that these have gold quality (using the monolingual approach, Section 4.2.) or near-gold quality (using the crosslingual approach, Section 4.4.). We constructed a total of nine emotion lexicons covering eight languages (including low-resource ones, such as Dutch and Indonesian). Table 5 depicts the number of lexical items for which we have generated previously unknown VAD or BE5 ratings per language. For illustration, we provide an analysis of the English BE5 lexicon (by far the largest resource constructed in this manner) in the remainder of this section.

Table 6 provides fundamental statistical characteristics of this newly developed data set. As can be seen, Joy ratings have higher mean, standard deviation and range than all the other categories. This suggests that a larger portion of lexical items expresses at least a moderate degree of Joy, whereas the other Basic Emotions are expressed less often and to a smaller extent. Table 7 lists the ten entries with the highest values for each Basic Emotion category. Obviously, the automatically derived ratings align well with our intuition, thus granting face validity to our approach.

Finally, Table 8 provides correlation values between the BE5 categories and VAD dimensions (ratings for the latter were taken from Warriner et al. (2013)). Joy displays a moderate negative correlation with the other Basic Emotions while these in turn have strong positive correlation among each other. Unsurprisingly, Valence displays a strong positive correlation with Joy and strong negative correlations with the remaining BE5 categories. Lastly, Arousal is uncorrelated with Joy but displays moderate positive correlation with Anger, Sadness, Fear and Disgust. These findings are consistent with empirically determined emotion values, thus validating our claims concerning the good quality of the constructed resources (Wierzba et al., 2015; Hinojosa et al., 2016a).

6. Conclusion

Progress in emotion analysis is hampered by a multitude of heterogeneous and, in the end, mutually incompatible emotion representation formats. In this paper, we performed the first large-scale analysis of *representation map-*

ping as a means to mediate between these heterogeneous formats. Our simple, yet highly effective, supervised approach makes use of the wide range of emotion lexicons already developed in various psychology labs.

We could show that, in the monolingual setup, automatic representation mapping outperforms human inter-study reliability and, therefore, produces *gold quality* data. In the crosslingual set-up, our approach still performs comparable to manual annotation though less robust than in the first set-up for some affective dimensions or categories, thus rendering *near-gold quality* entries. In both set-ups, mapping existing ratings to another format performs way better than the state-of-the-art in emotion prediction. Hence, we conjecture that our approach paves the way to greatly improve interoperability and re-use of lexical resources in this field. Lastly, we applied our technique to produce (near) gold quality emotion lexicons for eight typologically diverse languages, including low-resourced ones. These resources (together with our code) are available via `github.com/JULIELab/EmoMap`.

7. Bibliographical References

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

Bradley, M. M. and Lang, P. J. (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.

Briesemeister, B. B., Kuchinke, L., and Jacobs, A. M. (2011). Discrete Emotion Norms for Nouns: Berlin Affective Word List (DENN–BAWL). *Behavior Research Methods*, 43(2):441.

Buechel, S. and Hahn, U. (2016). Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 1114–1122.

Buechel, S. and Hahn, U. (2017a). A flexible mapping scheme for discrete and dimensional emotion representations: Evidence from textual stimuli. In *CogSci 2017 — Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 180–185.

Buechel, S. and Hahn, U. (2017b). EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 578–585.

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

Cohen, P. R. (1995). *Emperical Methods for Artificial Intelligence*. MIT Press, Cambridge, MA.

Davidson, P. and Innes-Ker, Å. (2014). Valence and arousal norms for Swedish affective words. Technical Report Volume 14, No. 2, Lund University.

Desmet, B. and Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Eilola, T. M. and Havelka, J. (2010). Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods*, 42(1):134–140.

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

Ferré, P., Guasch, M., Martínez-García, N., Fraga, I., and Hinojosa, J. A. (2017). Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories. *Behavior Research Methods*, 49(3):1082–1094.

Guasch, M., Ferré, P., and Fraga, I. (2015). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, 48(4):1358–1369.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *ACL-EACL 1997 — Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics & 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

Hinojosa, J. A., Martínez-García, N., Villalba-García, C., Fernández-Folgueiras, U., Sánchez-Carmona, A., Pozo, M. A., and Montoro, P. R. (2016a). Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, 48(1):272–284.

Hinojosa, J. A., Rincón-Pérez, I., Romero-Ferreiro, M. V., Martínez-García, N., Villalba-García, C., Montoro, P. R., and Pozo, M. A. (2016b). The Madrid Affective Database for Spanish (MADS): Ratings of dominance, familiarity, subjective age of acquisition and sensory experience. *PLoS One*, 11(5):e0155866.

Imbir, K. K. (2016). Affective Norms for 4900 Polish Words Reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Frontiers in Psychology*, 7:#1081.

Kanske, P. and Kotz, S. A. (2010). Leipzig affective norms for German: A reliability study. *Behavior Research Methods*, 42(4):987–991.

Li, S., Xu, J., Zhang, D., and Zhou, G. (2016). Two-view label propagation to semi-supervised reader emotion classification. In *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics*, volume Technical Papers, pages 2647–2655.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, New York, NY.

Monnier, C. and Syssau, A. (2014). Affective norms

for French words (FAN). *Behavior Research Methods*, 46(4):1128–1137.

Montefinese, M., Ambrosini, E., Fairfield, B., and Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46(3):887–903.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., and Brysbært, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45(1):169–177.

Palogiannidi, E., Koutsakis, P., Iosif, E., and Potamianos, A. (2016). Affective lexicon creation for the Greek language. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2867–2872.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP 2002 — Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Pinheiro, A. P., Dias, M., Pedrosa, J., and Soares, A. P. (2017). Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behavior Research Methods*, 49(2):698–716.

Redondo, J., Fraga, I., Padrón, I., and Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

Ric, F., Alexopoulos, T., Muller, D., and Aubé, B. (2013). Emotional norms for 524 French personality trait words. *Behavior Research Methods*, 45(2):414–421.

Riegel, M., Wierzba, M., Wypych, M., Żurawski, Ł., Jednoróg, K., Grabowska, A., and Marchewka, A. (2015). Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL–R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

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

Schmidtke, D. S., Schröder, T., Jacobs, A. M., and Conrad, M. (2014). ANGST: Affective Norms for German Sentiment Terms, derived from the affective norms for English words. *Behavior Research Methods*, 46(4):1108–1118.

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

Sedoc, J., Preoţiuc-Pietro, D., and Ungar, L. H. (2017). Predicting emotional word ratings using distributional representations and signed clustering. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, pages 564–571.

Sianipar, A., van Groenestijn, P., and Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Frontiers in Psychology*, 7:#1907.

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., and Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior research methods*, 44(1):256–269.

Sobhani, P., Mohammad, S. M., and Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In **SEM 2016 — Proceedings of the 5th Joint Conference on Lexical and Computational Semantics @ ACL 2016*, pages 159–169.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Stadthagen-Gonzalez, H., Imbault, C., Pérez-Sánchez, M. A., and Brysbært, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1):111–123.

Stevenson, R. A., Mikels, J. A., and James, T. W. (2007). Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, 39(4):1020–1024.

Strapparava, C. and Mihalcea, R. (2007). SemEval 2007 Task 14: Affective text. In *SemEval 2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007*, pages 70–74.

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., and Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL–R). *Behavior Research Methods*, 41(2):534–538.

Wang, J., Yu, L.-C., Lai, K. R., and Zhang, X. (2016). Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957–1968.

Warriner, A. B., Kuperman, V., and Brysbært, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Wierzba, M., Riegel, M., Wypych, M., Jednoróg, K., Turnau, P., Grabowska, A., and Marchewka, A. (2015). Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLoS One*, 10(7):e0132305.

Yao, Z., Wu, J., Zhang, Y., and Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4):1374–1385.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. R., and Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545.