A Flexible Mapping Scheme for Discrete and Dimensional Emotion Representations: Evidence from Textual Stimuli

Sven Buechel (sven.buechel@uni-jena.de)

Udo Hahn (udo.hahn@uni-jena.de)

Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena,
Fürstengraben 27, D-07743 Jena, Germany

http://www.julielab.de

Abstract

While research on emotions has become one of the most productive areas at the intersection of cognitive science, artificial intelligence and natural language processing, the diversity and incommensurability of emotion models seriously hampers progress in the field. We here propose kNN regression as a simple, yet effective method for computationally mapping between two major strands of emotion representations, namely dimensional and discrete emotion models. In a series of machine learning experiments on data sets of textual stimuli we gather evidence that this approach reaches a human level of reliability using a relatively small number of data points only.

Keywords: Models of Human Emotion; Representation Mapping; Machine Learning; Natural Language Processing

Introduction

In the past decades, a multitude of different models have been devised to elucidate the nature of human emotion (Scherer, 2000). A common distinction at the representational level of emotions sets *dimensional* models apart from *discrete* or *categorical* models (Stevenson, Mikels, & James, 2007).

Dimensional models consider affective states to be best described relative to a small number of independent emotional *dimensions* (often two or three). Substantial contributions to this line of research are often attributed to Osgood, Suci, and Tannenbaum (1957) as well as Mehrabian and Russell (1974) (Scherer, 2000). Although different labels have been proposed by major proponents of this approach, we here refer to these fundamental dimensions as *Valence* (the positiveness or negativeness of an emotion), *Arousal* (a calm–excited scale) and *Dominance* (the perceived degree of control over a (social) situation)—*VAD*, in short.[1]

Discrete models, on the other hand, often refer to emotions as evolutionary derived response pattern to major environmental events—each with its specific elicitation conditions (Scherer, 2000). Thus, in contrast to dimensional models which tend to focus on the subjective feeling aspect of emotion (and its associated verbal expression) researchers who adhere to the discrete approach rather tend to focus on motor (especially facial) expression and adaptive behavior. Among others, Plutchik (1980), Izard (1994) and Ekman (1992) are most influential for the development of this line of research. Although many different sets of such *basic emotions* have been proposed (typically ranging between 7 and 14 categories), up until now, no consensus has been reached on their exact and complete number (Scherer, 2000). However, most researchers seem to agree on at least five basic categories, namely *Joy*, *Anger*, *Sadness*, *Fear*, and *Disgust*.

For dimensional models, a broad variety of stimulus data bases have been developed, predominantly covering lexical stimuli. The *Affective Norms for English Words* (ANEW) (Bradley & Lang, 1999a) have been one of the first and probably most important data sets which comprise affective norms for Valence, Arousal and Dominance for 1,034 English words. Complementary lexical affective norms have also been developed for a wide range of other languages, such as German, Spanish or Polish (Võ et al., 2009; Redondo, Fraga, Padrón, & Comesaña, 2007; Riegel et al., 2015). In addition, larger linguistic units have been considered for emotion assessment moving ratings from lexical items up to sentence and text level (Pinheiro, Dias, Pedrosa, & Soares, 2017; Bradley & Lang, 2007), on the one hand, and considering alternative modalities, such as pictures and sounds, on the other hand (Lang, Bradley, & Cuthbert, 2008; Bradley & Lang, 1999b). Although these stimulus sets were primarily created for dimensional representations, research activities increasingly covered discrete emotion representations, as well (for all modalities). Consequently, many of the stimuli which have formerly been rated according to affective dimensions only, in the meantime, have also received discrete categorical norm ratings in terms of double encodings (e.g., Stevenson and James (2008), Stevenson et al. (2007) and Libkuman, Otani, Kern, Viger, and Novak (2007); see Table 1 for a list of resources with both dimensional and discrete ratings).

These resources have been highly influential for artificial intelligence (AI) research: Within the broader context of affective computing (Picard, 1997), they have specifically fostered the prediction of affective states from textual stimuli which is—as a subtask of natural language processing (NLP)—most commonly referred to as *sentiment analysis* (Pang & Lee, 2008; Liu, 2015; Mohammad, 2016). At the outset, NLP researchers focused on the Valence dimension only, typically trying to assign a piece of text to either a positive or a negative class (Pang, Lee, & Vaithyanathan, 2002). In the meantime, the interest in more advanced

[1] Another common name for the *Valence* dimension is *Pleasure* (PAD). Our choice of terminology (VAD) follows the more recent stimulus sets we use here (Warriner, Kuperman, & Brysbært, 2013; Ferré, Guasch, Martínez-García, Fraga, & Hinojosa, 2016).

models of emotions (going beyond positive-negative polarity judgments) has increased considerably. At first, this development was centered around discrete models (Ovesdotter Alm, Roth, & Sproat, 2005; Strapparava & Mihalcea, 2007), whereas only very recently the interest in dimensional models rapidly began to rise, as well (Buechel & Hahn, 2016; Wang, Yu, Lai, & Zhang, 2016; Sedoc, Preoţiuc-Pietro, & Ungar, 2017)—a focal change that profoundly benefited from the availability of affective norms developed in psychology labs. Ironically, in NLP, we now face a situation where the enormous interest in analyzing affectively loaded language has led to a proliferation of competing formal representation schemes for affective states whose motivation can be traced in various branches of psychological emotion theory (Valence-only, Valence-Arousal-Dominance, different sets of basic emotions, etc.). Consequently, it has become increasingly difficult to reliably compare the performance of different emotion recognition algorithms (Buechel & Hahn, 2016).

A possible solution to this dilemma is to elaborate explicit mappings between different representation formats, i.e., to predict the affective norm of a stimulus according to *one representation format* when the norm is already known in *another format* (e.g., dimensional and discrete representations; see Figure 1 for a graphical illustration). Not only would this affect formerly incommensurable algorithms but also widely ease the reusability of text collections annotated with different emotional ratings—one of the most important factors for advances given the predominance of training data-dependent supervised machine learning in NLP. In fact, not only computationally focused research would benefit from such mappings but also empirical research in psychology and cognitive science (Stevenson et al., 2007). By that, both the dimensional and the discrete view on emotion would be further integrated so that empirical findings from one view (e.g., regarding priming or memory) could be more directly compared to findings from the other view. Furthermore, existing stimulus sets originally based on one of these approaches could be easily enriched by norms employing other encoding schemes so that researchers could choose from a number of alternative, though mutually translatable emotion representation formats when designing experiments. This outlook becomes even more promising when we take into account the vast number of stimuli sets which bear ratings according to dimensional *and* discrete formats (see Table 1).

Despite the benefits of transferability, previous work on *automatically* translating between those formats (in contrast to manual re-annotation) has been relatively rare in the fields of psychology and AI. Stevenson et al. (2007) collected discrete ratings in addition to the dimensional ratings of ANEW. They repeated this effort in a follow-up study for the *International Affective Digitized Sounds* (IADS) stimulus set (Stevenson & James, 2008; Bradley & Lang, 1999b). Performing multiple linear regression between the categories/dimensions of both formats they evaluated the predictive power of the elements of the source representation (dimensions or categories) by the

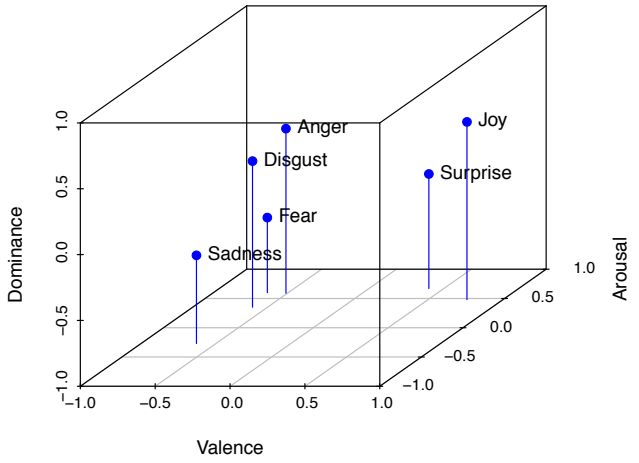

Figure 1: Affective space spanned by the Valence-Arousal-Dominance model, together with the position of six basic emotions (as determined by Russell and Mehrabian (1977); figure adapted from Buechel and Hahn (2016)).

statistical significance of their β-coefficients. They conclude that neither any of the affective dimensions consistently predict (one of the) discrete categories nor can predictions be made the other way round. The findings of Pinheiro et al. (2017) on their own data set of Portuguese sentences, in principle, support this conclusion. The present study differs from these precursors by concentrating on the combined model performance (and not on the contribution of the individual independent variables to it).

In contrast with these rather negative interpretations, in AI research, such emotion mappings have already been implemented with quite promising results. Calvo and Kim (2013) presented an algorithm that determines the emotional category of a text based on dimensional word ratings from psychology, using VAD as an interim representation before mapping onto discrete categories. Similarly, Buechel and Hahn (2016) presented a tool for predicting VAD scores from texts which maps their output onto basic emotions using support vector machines. Not only did they achieve highly competitive results regarding their emotion predictions, but they also report on a surprisingly high mapping performance (up to $R^2 = .944$ when predicting Valence given numerical scores for five basic emotions).

In this contribution, we follow up on this line of research by presenting a series of machine learning experiments that scrutinize the capability of such mapping schemes for textual stimuli. We restrict ourselves to well-known data sets of relatively small size so that the implications of our work can be put into practice without further restrictions (e.g., data limitations in a specific domain). For modeling, we decided to rely on k Nearest Neighbor (kNN) regression because of its simplicity, thus demonstrating that even elementary machine learning methods are sufficient here.

Our experiments fall into three steps. First, we generally demonstrate the feasibility of our approach by examining

the mapping performance between discrete and dimensional emotion formats on two different data sets, an English and a Spanish one. Second, we investigate how well these models generalize over different data sets and languages. In a third step, we examine how well this approach can be ported from psychology to NLP.

Study A: Mapping within a Stimulus Set

In the first experiment, we examine the capability of machine learning techniques to map dimensional and discrete emotion formats onto each other when training and test data are derived from the same data set.

Method

Material. We compose two different stimulus sets each receiving dimensional and discrete ratings from individual contributions. The first data set is ANEW which carries norms for Valence, Arousal and Dominance as supplied by Bradley and Lang (1999a); later on Stevenson et al. (2007) added discrete norms for Joy, Anger, Sadness, Fear and Disgust to it. The first half of the second set was originally presented by Redondo et al. (2007) as the Spanish adaptation of ANEW, thus including direct Spanish translations from the original English items. 1,012 of these words overlap with the ones rated by Ferré et al. (2016) according to basic emotions (together forming the second stimulus set). For both the English and the Spanish stimulus set, dimensional ratings were assigned using a 9-point SAM (a set of human-like pictograms displaying different levels of Valence, Arousal and Dominance (Bradley & Lang, 1994)). For the emotional categories, 5-point scales ranging from *not at all* to *extremely* were used. We use mean ratings by all subjects as supplied by the respective authors without performing any further transformation of the data (e.g., re-scaling).

Procedure. We used the R package CARET[2] to train kNN models in order to map between dimensional and discrete emotion representation schemes. For each dimension/category of the target representation, an individual model was trained given all the dimensions/categories of the source representation as features (e.g., there is *one* model to predict Anger given Valence, Arousal *and* Dominance ratings as input). We ran a 10-fold cross-validation (90% of the data were used for training and hyper-parameter tuning and the remaining 10% were made available for testing; the process was repeated ten times averaging the results). For the hyper-parameter k a grid search was performed repeating the procedure for each integer in the interval $[1,100]$. Consequently, the k-values may vary across the individual models. For comparability between different contributions, Pearson's r was used to assess the goodness of the fit.

Results

Table 2, section "Study A", depicts the results of the cross-validation (data sets in rows, *target* dimension/category in columns). As can be seen, the results range roughly between $r \approx .73$ up to .97 (both for mapping onto VAD on the English data set). We consider these figures to be surprisingly high, given the small amount of data points we have (from a machine learning point of view) and the elementary model we chose. Henceforth, for comparing correlation coefficients, we use two-tailed Z-tests for independent samples (tests for *dependent* samples are not eligible due to our cross-validation methodology). We find that mapping from dimensional to discrete ratings performs significantly better on the English data set than mapping the other way round ($z = 2.42$, $p < .05$), while the difference in mapping accuracy is not significant regarding the Spanish data ($z = 0.74$, $p \geq .05$).

Next we compare our model's fit against human reliability. Warriner et al. (2013) replicated the ratings of ANEW finding a correlation of their novel data with the original norms of $r = .953$, .759 and .795 for Valence, Arousal and Dominance, respectively. Thus, on the English data set, computationally mapping discrete emotion norms to dimensional ratings results in a significantly higher correlation with the original values than this replication study regarding Valence and Dominance (Valence: $z = 4.1$, $p < .001$; Dominance: $z = 3.1$, $p < .001$). For Arousal, the results are not significantly different ($z = 1.72$, $p \geq .05$).

Study B: Crosslingual Mapping

In our second experiment, we examine in how far the above models generalize over different studies and languages.

Method

We use the same English and Spanish data sets as for the previous experiments (with both dimensional and discrete ratings). In line with Study A, we train an individual model for each target category/dimension using all dimensions or categories of the source format (discrete or dimensional) as features. The k-parameter was chosen according to the highest performance in the 10-fold cross-validation set-up from the prior experiment. This time, the models were trained *on the whole* of one data set and then mutually tested on the other one (so that eight models are trained on the English data—one for each dimension/category—and then tested on the Spanish data, and the other way round). Therefore, no cross-validation is necessary. Performance is measured as correlation between predicted and actual values.

Results

Overall, we find that the models trained on the English data generalize well over the Spanish data and vice versa (see Table 2, Study B). The drops in performance (compared to Study A) regarding the individual dimensions/categories range well below 10% points. In fact, regarding the average performance of mapping basic emotions onto VAD dimensions, for neither of the two data sets the correlation decreases significantly (comparison relative to the target data; mapping *to* English: $z = 1.46$, $p \geq .05$, *to* Spanish: $z = 1.6$, $p \geq .05$). Regarding the mapping from the dimensional to the discrete

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

Table 1: Overview of selected stimulus sets bearing ratings according to both dimensional and discrete models.

Stimuli	Overlap	Dimensional Ratings	Discrete Ratings
words	1,012	Redondo et al. (2007)	Ferré et al. (2016)
	1,036	Bradley and Lang (1999a)	Stevenson et al. (2007)
sentences	1,192	Buechel and Hahn (2017)	Strapparava and Mihalcea (2007)
	192	Pinheiro et al. (2017)	Pinheiro et al. (2017)
images	703	Lang et al. (2008)	Libkuman et al. (2007)
sounds	111	Bradley and Lang (1999b)	Stevenson and James (2008)

Table 2: Results for studies A, B and C in Pearson's *r* relative to the data sets on which the models are trained and tested on (English, Spanish and EMOBANK (EMOB.)), and what the input and what the target emotion format for the mapping is (dimensional or discrete). **Av**: Average over the respective correlation coefficients (dimensional (VAD) or discrete basic emotions (BE)).

Study	Data	Dimensional→Discrete						Discrete→Dimensional			
		Joy	Ang.	Sad.	Fear	Dsg.	$\mathbf{Av_{BE}}$	Val.	Aro.	Dom.	$\mathbf{Av_{VAD}}$
A	English→English	0.960	0.873	0.863	0.868	0.798	**0.872**	0.967	0.725	0.840	**0.844**
	Spanish→Spanish	0.959	0.848	0.826	0.872	0.743	**0.849**	0.971	0.743	0.860	**0.858**
B	English→Spanish	0.948	0.791	0.807	0.829	0.698	**0.815**	0.966	0.740	0.808	**0.838**
	Spanish→English	0.948	0.831	0.855	0.841	0.772	**0.850**	0.963	0.715	0.795	**0.825**
C	EMOB.→EMOB.	0.738	0.481	0.674	0.559	0.348	**0.560**	0.788	0.227	0.412	**0.476**
	English→EMOB.	0.643	0.411	0.637	0.518	0.301	**0.502**	0.682	0.156	0.360	**0.400**
	Inter-Rater Reliab. EMOB.	0.599	0.495	0.682	0.638	0.445	**0.572**	–	–	–	–

format, losses in performance are significant, however, only by a small margin when mapping from Spanish to English ($z = 1.98$, $p < .05$; to Spanish: $z = 2.52$, $p < .05$). Comparing our models to human reliability (see above), we find that, for Valence, the predictions by the models trained on the Spanish data have still a significantly higher correlation with the original norms by Bradley and Lang (1999a) than the reproduced norms by Warriner et al. (2013) ($z = 2.81$, $p < .001$). For Arousal, the reproduction yields a significantly higher correlation ($z = 2.19$, $p < .5$) while for Dominance the difference is not significant ($z = 0.03$, $p \geq .05$).

Study C: Application to NLP Data Set

In the third experiment, we examine whether the mapping approach from the previous two studies translates to a concrete NLP scenario, given the task to automatically enrich existing emotion data sets with complementary emotion formats.

Method

We here rely on the recently developed EMOBANK data set[3] (Buechel & Hahn, 2017) which comprises 10k sentences together with their VAD ratings. To the best of our knowledge, EMOBANK is the only NLP resource annotated for multiple emotion formats: A subset of 1,192 sentences (English news headlines) has formerly been annotated for six emotion categories on a $[0, 100]$ scale by Strapparava and Mihalcea (2007). We use this subset, first, to train kNN models in a cross-validation set-up (as in study A), and second, to evaluate the performance of the models previously trained on the English stimulus set on these novel ratings (as in Study B).

Results

This set-up yields three main results. First, the overall mapping performance drops sharply compared to the former two studies. Comparing the cross-validation performance of our models from the English stimulus set (Study A) with those of the EMOBANK data (Table 2, Study C), we find a considerable decrease in correlation of about 35 percentage points (comparing average correlation coefficients for basic emotions and VAD; $z = 15.99$ and 16.21, respectively, $p < .001$).

In contrast to these mediocre results, the second main finding can be summarized such that our performance does only decrease by a small margin when the models are not trained on EMOBANK but on the English stimuli from Study A (comprising words instead of headlines and gathered with a dissimilar methodology; first *vs.* second line of Table 2, Study C). For mapping onto VAD, the drop is still statistically significant ($z = 2.11$, $p < .05$) while for mapping onto BE it is not ($z = 1.82$, $p \geq .05$). This suggests that, although our approach works better for lexical data gathered in psychological settings than for headlines annotated in NLP frameworks, the models still generalize well in the sense that one can apply models trained on the former to the latter without sacrificing a lot of performance.

[3] https://github.com/JULIELab/EmoBank

Even more surprisingly, our third main finding is that our approach still performs very well compared to human reliability (see bottom row of Table 2). Inter-rater reliability is reported by Strapparava and Mihalcea (2007) as the correlation of *one* rater with the mean judgment of the *remaining* raters averaged over *all* raters. Therefore, the output of our models can be cautiously compared against these reliability values. In this setting, we find no significant difference regarding the average over the basic emotions ($z = 0.4$, $p \geq .05$). We carefully interpret this observation to indicate that our output correlates with the aggregated rating of several subjects about as good as an average human does. Thus, consistent with our findings from Study A and B, our approach appears to perform comparably to human subjects and, in fact, even predicts normative Joy ratings significantly better ($z = 6.02$, $p < .001$). This suggests that the performance drop highlighted as the first main finding might point at different levels of data quality rather than taking this as evidence that our approach might be unsuitable for NLP data (we will get back to this issue in the subsequent discussion section).

General Discussion

We presented a series of experiments in which we examined the level of performance that can be achieved for mapping emotion ratings onto each other following the dimensional or the discrete representation format for the case of textual stimuli. To make our work more informative in terms of immediate reusability, we limited ourselves to employing relatively small and commonly used data sets, as well as elementary machine learning techniques.

In study A, we took into account two data sets from psychology, an English and a Spanish word stimulus set, each one bearing dimensional *and* discrete emotion ratings. On both sets, the mapping performance was surprisingly high. When comparing our prediction accuracy to a reassessment study of the English norms with human subjects, we found that our predicted values yielded significantly higher correlation with the original ratings than the novel reproduction regarding two of the three VAD dimension. This astonishing result suggests that given affective ratings in one format, ratings for the complementary emotion format can be computationally induced at a human level of reliability.

Study B goes beyond these considerations by asking how well these models generalize over different data sets with focus on different languages. The observation that the decrease in average mapping performance is only statistically significant in half of the cases suggests that the models generalize well over different (European) languages. However, it must be taken into account that the English and Spanish data sets are direct translations of each other regarding their raw data, possibly boosting the pairwise reusability of the models.

In Study C, we investigated a realistic usage scenario for our approach. Instead of lab data sets typically used in psychology, we here focused on a recently developed corpus of real-world news headlines, again annotated for both emotional dimensions and categories. This set-up yielded three results. First, compared to the former studies, we found a strong decrease in overall mapping performance. Second, the difference between the models directly trained on these data and the ones transfered from Study A were quite small (not even significant for mapping onto BEs). And third, our data suggest that our approach is on par with human annotation performance, despite the overall drop in mapping accuracy.

A possible explanation for this somewhat inconsistent behavior could be that, while the psychological data sets consist of word stimuli with explicit selection criteria, EMOBANK comprises "real-world" language data (news headlines instead of individual words). Thus, subjects can interpret these stimuli in a greater number of ways and may also be more strongly affected by biases from, e.g., political orientation or personal biography. In addition, the stimuli from Studies A and B have typically received a greater number of individual ratings which makes their aggregation potentially more reliable (i.e., less *noisy* in terms of training data).

Besides the above considerations, the results from Study C actually support the flexibility of the approach outlined here. Especially the observation that our models for *mapping existing annotations* operate about as accurately as a single human rater *freshly annotating new raw data* suggests that soon we may be able to fully automatically translate affective norms in terms of VAD to basic emotions and vice versa.

In conclusion, the experiments we presented here clearly demonstrate the power as well as the possible impact of our (still rather simple) set-up. The perspective of being reliably able to map back and forth between those popular emotion formats could not only lead to an improved availability of emotionally rated data sets in psychology and NLP. In addition, it may promote the integration of both views on emotion in psychological theory. Despite only presenting evidence from textual stimuli, we suggest that our approach may work for other modalities (and possibly across modalities) as well because no linguistic information was used for the prediction. We will address this conjecture in future work.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful suggestions and comments.

References

Bradley, M. M., & 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., & Lang, P. J. (1999a). *Affective Norms for English Words (*ANEW*): Stimuli, instruction manual and affective ratings* (Tech. Rep. No. C-1). Gainesville, FL: University of Florida.

Bradley, M. M., & Lang, P. J. (1999b). *International Affective Digitized Sounds (*IADS*): Stimuli, instruction manual and affective ratings* (Tech. Rep. No. B-2). Gainesville, FL: University of Florida.

Bradley, M. M., & Lang, P. J. (2007). *Affective Norms for English Text (*ANET*): Affective ratings of text and instruction manual* (Tech. Rep. No. D-1). Gainesville, FL: University of Florida.

Buechel, S., & 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* (pp. 1114–1122).

Buechel, S., & Hahn, U. (2017). 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 ACL* (Vol. 2: Short Papers, pp. 578–585).

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

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

Ferré, P., Guasch, M., Martínez-García, N., Fraga, I., & Hinojosa, J. A. (2016). Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories. *Behavior Research Methods*. (Online First Article) doi: 10.3758/s13428-016-0768-3

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*(2), 288–299.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International Affective Picture System (*IAPS*): Affective ratings of pictures and instruction manual* (Tech. Rep. No. A-8). Gainesville, FL: University of Florida.

Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., & Novak, N. (2007). Multidimensional normative ratings for the International Affective Picture System. *Behavior Research Methods*, *39*(2), 326–334.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. New York, NY: Cambridge U.P.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion Measurement* (pp. 201–237). Oxford, U.K.: Elsevier.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: Univ. of Illinois Press.

Ovesdotter Alm, E., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *HLT-EMNLP 2005 — Proc. of the Human Language Technology Conference & Conference on Empirical Methods in Natural Language Processing* (pp. 579–586).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

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

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Pinheiro, A. P., Dias, M., Pedrosa, J. a., & 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.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience* (Vol. 1: Theories of Emotion, pp. 3–33). New York, NY: Academic Press.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, *39*(3), 600–5.

Riegel, M., Wierzba, M., Wypych, M., Żurawski, L., Jednoróg, K., Grabowska, A., & 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., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273–294.

Scherer, K. R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), *The neuropsychology of emotion* (pp. 137–162). Oxford, U.K.: Oxford University Press.

Sedoc, J., Preoţiuc-Pietro, D., & 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 ACL* (Vol. 2: Short Papers, pp. 564–571).

Stevenson, R. A., & James, T. W. (2008). Affective auditory stimuli: Characterization of the International Affective Digitized Sounds (IADS) by discrete emotional categories. *Behavior Research Methods*, *40*(1), 315–321.

Stevenson, R. A., Mikels, J. A., & 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., & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *SemEval-2007 — Proc. of the 4th Intl. Workshop on Semantic Evaluations* (pp. 70–74).

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & 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., & Zhang, X. (2016). Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 2: Short Papers, pp. 225–230).

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