Predicting Emotional Word Ratings using Distributional Representations and Signed Clustering

João Sedoc and **Daniel Preoţiuc-Pietro** and **Lyle Ungar**
Positive Psychology Center
Computer & Information Science
University of Pennsylvania

Abstract

Inferring the emotional content of words is important for text-based sentiment analysis, dialogue systems and psycholinguistics, but word ratings are expensive to collect at scale and across languages or domains. We develop a method that automatically extends word-level ratings to unrated words using signed clustering of vector space word representations along with affect ratings. We use our method to determine a word's valence and arousal, which determine its position on the circumplex model of affect, the most popular dimensional model of emotion. Our method achieves superior out-of-sample word rating prediction on both affective dimensions across three different languages when compared to state-of-the-art word similarity based methods. Our method can assist building word ratings for new languages and improve downstream tasks such as sentiment analysis and emotion detection.

1 Introduction

Word-level ratings play an important role in computational linguistics and psychology research. Many studies have focused on collecting ratings related to the properties of words, such as frequency, complexity, concreteness, imagery, age of acquisition, familiarity and affective states (Kuperman et al., 2012; Schock et al., 2012; Juhasz and Yap, 2013; Brysbaert et al., 2014). Applications span from memory experiments to developing reading tests and analyzing texts from non-native speakers (Mohammad and Turney, 2013). In NLP, these ratings can be used to quantify different properties in large scale naturally occurring text, for example when analysing lexical choice between demographic groups (Preoţiuc-Pietro et al., 2016b) or music lyrics (Maulidyani and Manurung, 2015).

Of particular importance to NLP research are ratings of affect, which can be used for sentiment analysis and emotion detection (Pang and Lee, 2008; Preoţiuc-Pietro et al., 2016a). The main dimensional model of affect is the circumplex model of Russell (1980), which posits that all affective states are represented as a linear combination of two independent systems: valence (or sentiment) and arousal (Posner et al., 2005). For example, the word 'fear' is rated by humans as low in valence (2.93/9) but relatively high in arousal (6.41/9), while the word 'sad' is low in both valence (2.1/9) and arousal (3.49/9).

However, collecting word ratings is very time consuming and expensive for new languages, domains or properties, which hinders their applicability and reliability. In addition, although word ratings are performed using anchoring to control for differences between raters, implicit biases may exist when rating. This can be caused by certain demographic biases or halo effects e.g., a high valence word is more likely to be rated higher in arousal. An independent way of measuring words could also help refine existing ratings, rather than only extending them to unrated words.

Automatically expanding affective word ratings has been studied based on the intuition that words similar in a reduced semantic space will have similar ratings (Recchia and Louwerse, 2015; Palogiannidi et al., 2015; Vankrunkelsven et al., 2015; Köper and Im Walde, 2016). For example, Bestgen and Vincze (2012) compute the rating of an unknown word as the average of its k-nearest neighbors from the low-dimensional semantic space. However, the downside is that antonyms are also semantically similar, which is expected to reduce the accuracy of these methods. Orthographic similarity has shown to slightly improve results (Recchia and Louwerse, 2015). A different approach to rating prediction is based on graph methods inspired by label propagation (Wang et al., 2016). In a related task of adjective intensity prediction, Sharma et al. (2015) also use distributional methods, but their work is restricted to discrete categories and relative ranking within each semantic property. Another related task to affective norm prediction is building sentiment and polarity lexicons (Turney, 2002; Turney and Littman, 2003; Velikovich et al., 2010; Yih et al., 2012; Tang et al., 2014; Hamilton et al., 2016). However, polarity is assigned to words in order to determine if a text is subjective and its sentiment, which is slightly different to word-level affective norms e.g., 'sunshine' is an objective word (neural polarity), but has a positive affective rating.

Our approach builds upon recent work in learning word representations and enriches these by integrating a set of existing ratings. Including this information allows our method to differentiate between words that are semantically similar, but on opposite sides of the rating scale. Results show that our automatic word prediction approach obtains better results than competitive methods and demonstrates the benefits of introducing existing ratings on top of the underlying word representations. The superiority of our approach holds for both valence and arousal word ratings across three languages.

2 Data

Our gold standard data is represented by affective norms of words. The ratings are obtained by asking human coders to indicate the emotional reaction evoked by specific words on 9-point scales: valence (1–negative to 9–positive) and arousal (from 1–calm to 9–excited).

Originally, word ratings were computed using trained raters in a laboratory setup. The Affective Norms for English Words (Bradley and Lang, 1999) – ANEW – contained ratings for valence and arousal, as well as dominance for only 1034 English words. Similar norms were obtained for Spanish (Redondo et al., 2007). Recently, crowdsourcing was used to derive ratings for larger sets of words using the ANEW ratings for anchoring and validation. Warriner et al. (2013) computed valence, arousal, and dominance scores for 13,915 English lemmas. A similar methodology was used to obtain affective norms for Dutch – 4,300 words (Moors et al., 2013) – and Spanish – 14,031 words (Stadthagen-Gonzalez et al., 2016). In our experiments, we use valence and arousal ratings for these three languages. Although some affective norms contain a third dimension of dominance (from feeling dominated to feeling dominant), we choose not to include this as it was not present in all data sets.

3 Method

Our method consists of two separate steps. First, we leverage large corpora of naturally occurring text and the distributional hypothesis in order to represent words in a semantic space with reduced dimensionality. Words that are similar in this space will appear in similar contexts, hence are expected to have similar scores. However, words of opposite polarity have similar distributional properties and will also be very similar in this space (Landauer, 2002). Hence, we perform an additional second step which distorts the word representations, here implemented using signed spectral clustering.

3.1 Distributional Word Representations

Distributional word representations or word embeddings make use of the *distributional hypothesis* – a word is characterised by the company it keeps – to represent words as low dimensional numeric vectors using large text corpora (Harris, 1954; Firth, 1957).

We use the word2vec algorithm (Mikolov et al., 2013), without loss of generality, to generate word vectors as it is arguably the most popular model out of the variety of existing word representations. The word2vec embeddings for English and Spanish have 300 dimensions and are trained on the Gigaword corpora (Parker et al., 2011; Mendonca et al., 2011). For Dutch, we use the word2vec embeddings with 320 dimensions from Tulkens et al. (2016). All words in the embeddings have minimal tokenization, with no additional stemming or lowercasing. Our vocabulary consists of the words that have ratings on either scale.

3.2 Signed Spectral Clustering

To infer the score of an unrated word we use a clustering approach – rather than nearest neighbors – to automatically uncover the number of related words based on which the rating is com-

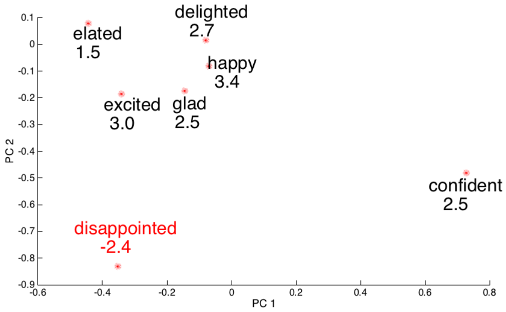

Figure 1: A continuous two-dimensional representation of a cluster (using K-means) of English words and their normalized valence ratings. After incorporating valence ratings using the signed clustering algorithm, “disappointed” is removed from the main cluster. The colors represent the resulting cluster memberships.

puted. Distributional word representations capture semantic word similarity. However, a common pitfall is that words with different properties can be used in similar contexts e.g., ‘happy’ and ‘sad’ are antonyms but are used similarly. Signed spectral clustering (SSC) – described in Sedoc et al. (2016) – is extremely well suited for this type of problem.

SSC is a multiclass optimization method which builds upon existing theory in spectral clustering (Shi and Malik, 2000; Yu and Shi, 2003; von Luxburg, 2007) and incorporates side information about word ratings in the form of negative edges which repel words with opposing scores from belonging to the same clusters. It minimizes the cumulative edge weights cut within clusters versus between clusters, while simultaneously minimizing the negative edge weights within the clusters.

More formally, given a partition of nodes of a graph into k clusters, $(A_1, \ldots, A_k)$, signed spectral clustering using normalized cuts minimizes

$$\sum_{j=1}^{k} \frac{\mathrm{cut}(A_j, \overline{A_j}) + 2\mathrm{links}^{-}(A_j, A_j)}{\mathrm{vol}(A_j)}.$$

For any subset A of the set of nodes, V, of the graph, let

$$\mathrm{vol}(A) = \sum_{v_i \in A} \sum_{j=1}^{|V|} |w_{ij}|,$$

where w_{ij} is the similarity or dissimilarity of words i and j. For any two subsets A and its complement $\overline{A}$, define

$$\mathrm{links}^{-}(A, A) = \sum_{\substack{v_i, v_j \in A \\ w_{ij} < 0}} -w_{ij}$$

$$\mathrm{cut}(A, \overline{A}) = \sum_{\substack{v_i \in A, v_j \in \overline{A} \\ w_{ij} \neq 0}} |w_{ij}|.$$

Note, that the main innovation of signed spectral clustering is minimizing the number of negative edges within the cluster, $\mathrm{links}^{-}(A_j, A_j)$. Without the addition of negative weights, signed spectral clustering is simply spectral clustering i.e., normalized cuts (Yu and Shi, 2003).

For this application, rather than incorporating a thesaurus knowledge base (a.k.a., side information) as in Sedoc et al. (2016), we used the continuous lexical scores from our arousal and valence ratings. To obtain signed information, we zero-centered the word ratings which are originally between 1 and 9. We create a similarity matrix where the weight between words i and j incorporate both the signed information and the word similarities computed using the cosine similarity of the distributional word representations. The similarity matrix W (a.k.a., weight matrix) is used to create word clusters which capture both the distributional features as well as the lexical features. We perform a separate clustering for each valence and arousal and each separate language. More formally, the similarity matrix

$$W = W^{emb} + \beta^{-} T^{-} \odot W^{emb} + \beta^{+} T^{+} \odot W^{emb}$$

where W^{emb} is the matrix of cosine similarities between vector embeddings of words, $\odot$ is element-wise multiplication. The matrix $T = T^{+} + T^{-}$ is the outer product of the normalized lexical ratings, where the matrices T^{+}, T^{-} contain the outer product of the normalized lexical ratings split into positive and negative entries, respectively, in matrix block form,

$$T^{+} = \begin{pmatrix} + & 0 \\ 0 & + \end{pmatrix}, T^{-} = \begin{pmatrix} 0 & - \\ - & 0 \end{pmatrix}.$$

The values β^{+} and β^{-} are found using grid search on the training data.

Figure 1 shows the intuition behind signed clustering by presenting an example cluster obtained using K-means clustering on the reduced semantic space (here showing the first two principal components). This includes the word ‘disappointed’ together with with words like ‘happy’, ‘excited’ and ‘elated’. While this is relatively appropriate for arousal, it is not the case for valence as they represent opposite ends of the rating spectrum. By incorporating valence information, ‘disappointed’ is taken apart from the cluster of words with positive valence and thus its negative valence rating will not be considered when predicting the rating of a word belonging to this cluster.

Note that we used signed spectral clustering (SSC) for our problem since, unlike when antonym pairs are used as side information, we need to incorporate continuous information. Other methods for adding antonym or arbitrary relationships on distributional word representations, are unable to extrapolate these to unseen words or handle unpaired side information (Yih et al., 2012; Chang et al., 2013; Faruqui et al., 2015; Mrkšić et al., 2016). Furthermore, our information comes in lists rather than sets, contexts, or patterns, which presents a problem for other existing methods (Tang et al., 2014; The Pham et al., 2015; Schwartz et al., 2015). An alternative to SSC – must-link / cannot-link clustering (Rangapuram and Hein, 2012) – has the downside of requiring a choice of threshold for defining the must-link and cannot-link underlying graph edges. An extended comparison of SSC to related methods is presented in (Sedoc et al., 2016).

4 Results

We compare the proposed method with other baselines and approaches which assign to the unrated word:

1. the mean of the available ratings (**Mean**);
2. the average of its k nearest rated neighbors in the semantic space – the method introduced in (Bestgen and Vincze, 2012) (**K-NN**);
3. the mean rating of words in its cluster using standard k-means clustering in the reduced semantic space (**K-Means**);
4. linear regression value with the word embedding dimensions as features (**Regression**);
5. the mean rating of words in its cluster using vanilla spectral clustering (i.e., $W = W^{emb}$) which uses normalized cuts (**NCut**), in order to measure the utility and impact of the signed spectral clustering.

We perform the experiment in a 10-fold cross-validation setup, where 90% of the ratings are known and used in training. Results are evaluated in both Root Mean Squared Error (RMSE) between the human and automatic rating and the Pearson Correlation Coefficient (ρ) between the list of human and automatic ratings. We used $k = 10$ nearest neighbors for **K-NN**, which generally outperforms $k = \{1, 5, 20\}$ over valence and arousal in all three test languages. This is consistent with the original results of Bestgen and Vincze (2012), although Recchia and Louwerse (2015) found that $k = 40$ was optimal for predicting arousal ratings. For all other clustering methods we used $k \sim 10\%$ of the total ratings ($k = 1000$ for English and Spanish, $k = 400$ for Dutch). In English valence experiments, the **K-means** cluster sizes have a median of 13 with $\sigma = 16.4$, for **NCut** the median is 6 with $\sigma = 62.5$ and for **SNCut** the median is 5 with $\sigma = 78.1$. In **SNCut**, smaller cluster sizes are associated with more extreme ratings.

The results are presented in Table 1 and show that our method (**SNCut**) consistently performs best across both ratings – valence and arousal – and across all three languages. For English and Spanish, the larger margins of improvement over the mean baseline and **K-NN** are obtained on valence. This is particularly intuitive, as opposite valence words are usually antonyms and are more useful to split apart compared to low/high arousal words, which might also not be as distributionally similar to each other. In all cases, the signed clustering step improves rating prediction significantly over vanilla spectral clustering (**NCut**), highlighting the utility of signed clustering. Out of the baseline methods, none consistently outperforms the others. In addition, we also used English 300 dimensional GloVe word embeddings (Pennington et al., 2014) instead of word2vec, which led to similar results using **SNCut** where for valence RMSE= 0.82, $\rho = 0.76$ and arousal RMSE= 0.73 and $\rho = 0.56$. As an upper bound comparison, Warriner et al. (2013) reported that the human inter-annotator agreements are 0.85 to 0.97, and 0.56 to 0.76 for valence and arousal respectively across various languages.

We also directly compare with results from previous work by matching the training and testing data sets where enough information was provided. When using only English ANEW words for out-of-sample analysis as in Recchia and Louwerse (2015), our results are slightly higher (ρ=.804 cf. ρ=.8 for valence, ρ=.632 cf ρ=.62 for arousal). We did not have enough information to reproduce

	English				Spanish				Dutch			
	Valence		Arousal		Valence		Arousal		Valence		Arousal	
Method	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ
Mean	1.274	0	0.896	0	1.331	0	0.930	0	1.050	0	0.842	0
K-NN (k=1)	1.265	0.533	1.048	0.308	1.328	0.011	1.359	0.012	0.977	0.409	0.976	0.407
K-NN (k=10)	0.961	0.659	0.764	0.523	1.035	0.644	0.862	0.465	0.949	0.557	0.727	0.544
K-Means	0.953	0.684	0.773	0.551	1.009	0.657	0.916	0.447	0.780	0.675	0.683	0.592
Regression	0.835	0.757	0.759	0.547	1.002	0.679	0.915	0.203	0.844	0.566	0.746	0.545
NCut	0.948	0.682	0.861	0.520	1.006	0.679	0.864	0.452	0.864	0.585	0.723	0.533
SNCut	**0.803**	**0.768**	**0.713**	**0.582**	**0.944**	**0.733**	**0.822**	**0.499**	**0.762**	**0.693**	**0.592**	**0.706**

Table 1: Accuracy of word rating prediction in a 10-fold cross-validation setup. For both English and Spanish the number of clusters for K-means, NCut and SNCut is 1000. For Dutch because of the reduced lexicon, we used 400 clusters.

their results on Spanish or Dutch, albeit their results (ρ=.52 valence and ρ=.36 arousal for Spanish; ρ=.50 valence and ρ=.47 arousal for Dutch) are far lower than our best results.

On the original 1,034 English ANEW ratings, Wang et al. (2016) used a 6:2:2 train/dev/test split and k-fold cross-validation. They achieve ρ=.801 for valence and ρ=.539 for arousal compared to ρ=.806 for valence and ρ=.615 for arousal when using our proposed method.

Figure 2 presents the rating prediction error of our method when varying the number of ratings used as seeds in signed clustering. As expected, the error of our predictions decreases with the amount of ratings available with signs of reaching a plateau towards the end.

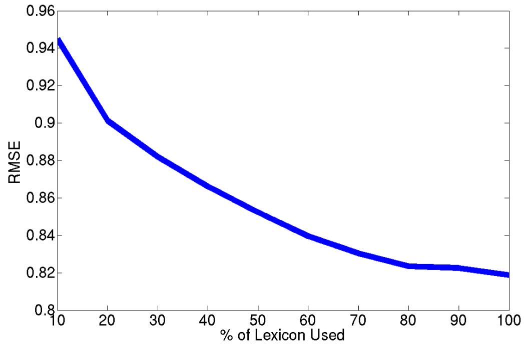

Figure 2: The RMSE of the signed clustering method (**SNCut**) as a function of the percentage of the lexicon ratings used for English valence prediction.

5 Conclusion

This study looked at the feasibility of automatically predicting word-level ratings – here valence and arousal – by combining distributional approaches with signed spectral clustering. Our experiments on word ratings of valence and arousal across three different languages showed that in an out-of-sample word rating prediction task, our proposed method consistently achieves the best prediction results when compared to a number of competitive methods and existing baselines.

Future work will include experiments on other word-level ratings, such as age-of-acquisition, dominance, imageability or abstractness, on other languages and using other word embeddings. Possible applications of our work include choosing the words to rate in an active learning setup on annotating new languages, automatically cleaning and checking word ratings and applying automatically derived scores to improve downstream tasks such as sentiment analysis or emotion detection.

Acknowledgments

The authors acknowledge the support of the Templeton Religion Trust, grant TRT-0048.

References

Yves Bestgen and Nadja Vincze. 2012. Checking and Bootstrapping Lexical Norms by Means of Word Similarity Indexes. *Behavior Research Methods*, 44(4):998–1006.

Margaret Bradley and Peter Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, Instruction Manual, and Affective Ratings. Technical report.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally known English Word Lemmas. *Behavior Research Methods*, 46(3):904–911.

Kai-Wei Chang, Wen-tau Yih, and Christopher Meek. 2013. Multi-relational latent semantic analysis. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1602–1612.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 1606–1615.

John R. Firth. 1957. A Synopsis of Linguistic Theory. In *Studies in Linguistic Analysis*, pages 1–32.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing Domain-specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 595–605.

Zelling Harris. 1954. Distributional Structure. *Word*, 10(23):146–162.

Barbara J Juhasz and Melvin J Yap. 2013. Sensory Experience Ratings for over 5,000 Mono-and Disyllabic Words. *Behavior Research Methods*, 45(1):160–168.

Maximilian Köper and Sabine Schulte Im Walde. 2016. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350000 German Lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC, pages 2595–2598.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition Ratings for 30,000 English Words. *Behavior Research Methods*, 44(4):978–990.

Thomas K Landauer. 2002. On the Computational Basis of Learning and Cognition: Arguments from LSA. *Psychology of Learning and Motivation*, 41:43–84.

Anggi Maulidyani and Ruli Manurung. 2015. Automatic Identification of Age-Appropriate Ratings of Song Lyrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 583–587.

Angelo Mendonca, Daniel Jaquette, David Graff, and Denise DiPersio. 2011. Spanish Gigaword Third Edition. *Linguistic Data Consortium*.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 746–751.

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

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of Valence, Arousal, Dominance, and Age of Acquisition for 4,300 Dutch Words. *Behavior Research Methods*, 45(1):169–177.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 142–148.

Elisavet Palogiannidi, E Losif, Polychronis Koutsakis, and Alexandros Potamianos. 2015. Valence, Arousal and Dominance Estimation for English, German, Greek, Portuguese and Spanish Lexica using Semantic Models. In *Proceedings of Interspeech*, Interspeech, pages 1527–1531.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. *Linguistic Data Consortium*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1532–1543.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology. *Development and Psychopathology*, 17(3):715–734.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elizabeth P. Shulman. 2016a. Modelling Valence and Arousal in Facebook Posts. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, NAACL, pages 9–15.

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016b. Discovering User Attribute Stylistic Differences via Paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI, pages 3030–3037.

Syama Sundar Rangapuram and Matthias Hein. 2012. Constrained 1-spectral clustering. *International conference on Artificial Intelligence and Statistics (AISTATS)*, 22:1143—1151.

Gabriel Recchia and Max M Louwerse. 2015. Reproducing Affective Norms with Lexical Co-occurrence

Statistics: Predicting Valence, Arousal, and Dominance. *The Quarterly Journal of Experimental Psychology*, 68(8):1584–1598.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish Adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Jocelyn Schock, Michael J Cortese, and Maya M Khanna. 2012. Imageability Estimates for 3,000 Disyllabic Words. *Behavior Research Methods*, 44(2):374–379.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric Pattern Based Word Embeddings for Improved Word Similarity Prediction. In *Proceedings of the 19th Conference on Computational Language Learning*, CoNLL, pages 258–267.

João Sedoc, Jean Gallier, Lyle Ungar, and Dean Foster. 2016. Semantic Word Clusters Using Signed Normalized Graph Cuts. *arXiv preprint arXiv:1601.05403*.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective Intensity and Sentiment Analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 2520–2526.

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel Perez Sanchez, and Marc Brysbaert. 2016. Norms of Valence and Arousal for 14,031 Spanish Words. *Behavior Research Methods*.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 1555–1565.

Nghia The Pham, Angeliki Lazaridou, and Marco Baroni. 2015. A Multitask Objective to Inject Lexical Contrast into Distributional Semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 21–26.

Stephan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC, pages 4130–4136.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Peter D Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL, pages 417–424.

Hendrik Vankrunkelsven, Steven Verheyen, Simon De Deyne, and Gerrit Storms. 2015. Predicting Lexical Norms using a Word Association Corpus. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 2463–2468.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The Viability of Web-derived Polarity Lexicons. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 777–785.

Ulrike von Luxburg. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 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.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Wen-tau Yih, Geoffrey Zweig, and John C Platt. 2012. Polarity Inducing Latent Semantic Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP, pages 1212–1222.

Stella X Yu and Jianbo Shi. 2003. Multiclass Spectral Clustering. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, ICCV, pages 313–319.