ANEW+: Automatic Expansion and Validation of Affective Norms of Words Lexicons in Multiple Languages

Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu and George Aaron Broadwell

State University of New York – University at Albany
1400 Washington Avenue Albany NY 12222
E-mail: sshaikh@albany.edu

Abstract

In this article we describe our method of automatically expanding an existing lexicon of words with affective valence scores. The automatic expansion process was done in English. In addition, we describe our procedure for automatically creating lexicons in languages where such resources may not previously exist. The foreign languages we discuss in this paper are Spanish, Russian and Farsi. We also describe the procedures to systematically validate our newly created resources. The main contributions of this work are: 1) A general method for expansion and creation of lexicons with scores of words on psychological constructs such as valence, arousal or dominance; and 2) a procedure for ensuring validity of the newly constructed resources.

Keywords: valence lexicon; expansion; validation

1. Introduction

Affective ratings of words are important to a broad spectrum of the language research community, including scholars engaged in sentiment and opinion analysis. The Affective Norms for English Words (ANEW) lexicon (Bradley and Lang, 2009) is a well-known and highly cited lexicon of words rated by human subjects along three dimensions – valence, arousal and dominance. The corpus consists of 2,477 words. The value for each word was obtained from ratings made by undergraduate students enrolled in a university. Ratings were made on a scale of 1 to 9, where a rating of 1 denoted highly negative and 9 denoted highly positive. As of December 2014, the ANEW corpus has received over 1,400 citations on Google Scholar, making it evident that it has a large impact in the scientific community. However, one major limitation of the ANEW corpus is the relatively small size. More recently, Warriner, Kuperman, and Brysbaert (2013) created a more extensive affective norms corpus, collecting ratings for approximately 14,000 words. Their normative procedure was highly similar to that used by Bradley and Lang's (2009). One noteworthy difference, however, is that Warriner et al.'s (2013) ratings were gathered using participants recruited through Mechanical Turk. Although the Warriner et al.'s (2013) paper is an important extension of the ANEW corpus, 14,000 words may not be sufficient for researchers who are working with a large amount of text, as is often the case in fields such as natural language processing. Thus, a larger corpus was needed. Furthermore, there is currently only one such resource (Redondo et al., 2007) for Spanish, and the number of words in this corpus is very limited (1,034); no comparable resources for affective norms of words currently exist for any other language.

In this article, we first describe our method of automatically expanding an existing affective lexicon for English. Our approach is general enough to apply to any specialized lexicon in any language where a partial resource exists. We also describe a method of automatically creating lexicons for a new language where no prior resources exist or are very limited; specifically, for Spanish, Russian and Farsi. We also describe the validation procedures used to ensure validity of these lexicons[1].

2. Related Work

There has been prior work in the automatic construction and expansion of affective lexicons using different techniques. Kim and Hovy (2004) used WordNet (Miller et al. 1995) to assign positive or negative polarity to words using synonyms and antonyms for a small set of seed words, however, such a method is limited by the set of seed words chosen. Esuli and Sebastini (2006) used semi-supervised learning to create SentiWordNet, where potentially every word in WordNet would be assigned a sentiment score, although many words actually may not be sentiment-bearing (cf. Taboada, 2011 for further discussion). A number of approaches use semantic proximity of words in variations of Latent Semantic Analysis (Turney and Littman, 2003; Bestgen, 2008; Bestgen and Vincze, 2012); however, their self-reported correlations of proposed expansions against human ratings are not sufficiently robust. Neilson (2011) created a new ANEW specifically geared towards detecting sentiment in microblog posts, but it only contains 2477 words scored manually on a scale of +2 (positive) to -2 (negative).

3. Approach

Our expansion method follows that adopted by Liu et al. (2014) for their automated expansion of the MRC psycholinguistic database. We use WordNet (Miller, 1995), a large English lexical database with over 150,000 words, hierarchically organized in synsets that capture semantically equivalent words. It is thus reasonable to assume that if one element of a synset has a known valence score, all other words in this synset should have the same or closely related scores, and can be added to the expanded lexicon with the inherited valence ratings.

[1] Available for download at the bottom of page: http://www.ils.albany.edu/research/projects/remnd/

3.1 Expansion of English Lexicon

The ANEW lexicon (Bradley & Lang, 2009) consists of 2,477 words, each assigned a valence score on a scale of 1 to 9, where a rating of 1 denotes highly negative and 9 denotes highly positive score. The Warriner et al. (2013) lexicon consists of such ratings for ~14,000 words.

As a first step towards creating an expanded English lexicon, we merged the ANEW and Warriner lexicons into one set of words. A correlation of *r*=0.953 was reported (Warriner et al., 2013) for values of the words shared between these two lexicons. Next, we used WordNet to impute the affect ratings for words that were derived from human raters to the words contained in their first (most frequent) synsets, i.e. synonyms of the most common meaning of the word. In Figures 1 and 2, we illustrate this process by examples. As shown in Figure 1, two synonyms of word *building* are *edifice* and *construction*. Thus, our expansion method will impute the affect value for source word *building* to both expansion words *edifice* and *construction*. In some cases (illustrated in Figure 2), multiple source words contribute to an expansion word, because the expansion word is in the first synset of different source words (e.g., the expansion word *atrocious* is in the first synset of the source words *horrible* and *awful*). In these cases, we take the average value of the source words to assign to the expansion word. Using this method, we expanded the combined set of ANEW and Warriner lexicons to the total of 22,756 words.

In another version of the expansion, we used all synonyms as well as hyponyms of source words to derive expansion words and impute their scores – resulting in an expanded set of 109,752 words. Other expansions can be similarly produced, using a subset of the WordNet synsets and hyponyms – e.g. using the top *N*-most synsets (sets of synonymous words) or other combinations.

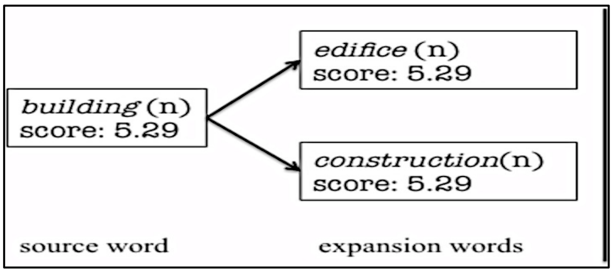

Figure 1. Expansion of source word building to two of its synonyms edifice and construction.

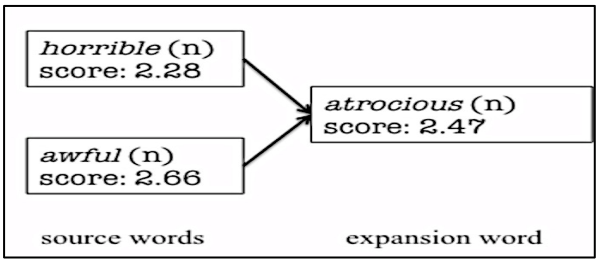

Figure 2. Expansion of source words horrible and awful to atrocious which is a synonym of both.

3.1 Validation of Expanded English Lexicon

We used two different approaches to provide converging evidence of the validity of our expansion method. The first approach is identical to that used by Warriner et al. (2013), which is to compare the affect values in our corpus to those obtained from a source that is established to be valid and reliable. The second approach compared the values imputed to the expansion words to ratings obtained using human subjects recruited through Amazon Mechanical Turk.

3.1.1. Validation against existing lexicons

Our first validation approach was to compare scores of expanded words against scores of words that we had human ratings for. To do so, we expanded only the words in ANEW lexicon and correlated the scores of expanded words if they overlapped with words in the Warriner lexicon. The observed correlation was Pearson's *r* = .661 for the words resulting from first synset expansion method, a highly statistically significant result, *p* < .001 (Figure 3). This suggests that the expansion method is indeed robust. The correlation for expansion words resulting from the all synonyms and hyponym expansion was *r*=0.57. This indicates that including more words through expansion reduces the correlation, although not below satisfactory levels. Since we wanted to use the most robustly correlated set for developing foreign language lexicons, we shall focus on the synonym expansion for the rest of this paper. The all synonym+hyponym expansion will be the focus of a separate publication.

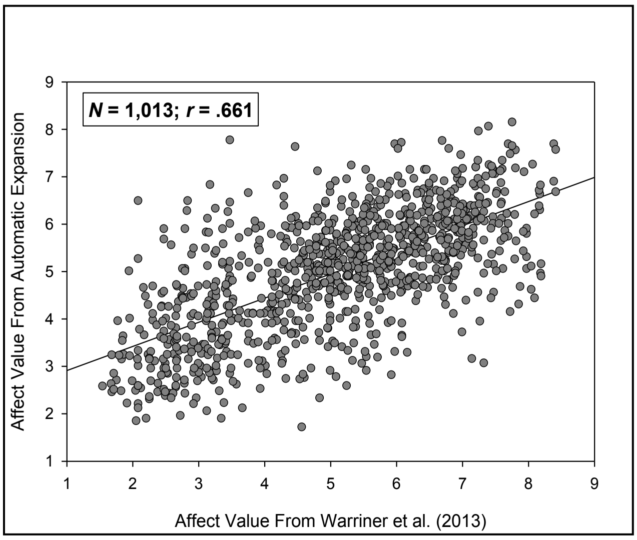

Figure 3. Scatterplot of the affect values of English words from our expansion method of the ANEW lexicon and those obtained by Warriner et al. (2013).

3.1.2. Validation using human judgments

As an additional validation step, we randomly selected 235 words from our expanded set of ~22K words, whose frequency of usage in the written text was variable (Log_HAL, taken from Balota et al., (2007) *Mean* = 5.17; *Standard Deviation* = 2.08; *Range* = .693-12.144). This selection was made to ensure that our random sample was representative of words that appear in various genres of written media, e.g., books, magazines etc. We also included 40 words randomly selected from the original Warriner lexicon. These 40 words and the 235 expansion words were presented as a single list in randomized order to workers on Amazon Mechanical Turk. To maximize the likelihood that our Turkers would provide high-quality data and take the task seriously: (1) the description of our study stated that we are looking for native English speakers, and (2) we imposed a restriction such that only Turkers who have completed at least 1,000 studies, with an approval rating of 99% were allowed to participate. The instructions provided to Turkers were similar to those provided by Bradley and Lang (2010) and Warriner et al. (2013) to their participants. Turkers had an unlimited amount of time to answer each word but could not return to a word once they have indicated their response. We collected data from 17 Turkers. To assess the reliability the ratings provided by turkers in our study, we first assessed the correlation of the affect values for the 40 words for which human gold standard already existed in the Warriner lexicon. The correlation for these words was nearly perfect, r = .96, giving us high confidence that the ratings obtained from turkers were reliable. The main analysis of interest was the correlation of the affect values for the 235 expansion words as derived automatically from our expansion method and those given by turkers. The correlation for this analysis was .759, a highly statistically significant result, p < .001 (Figure 4).

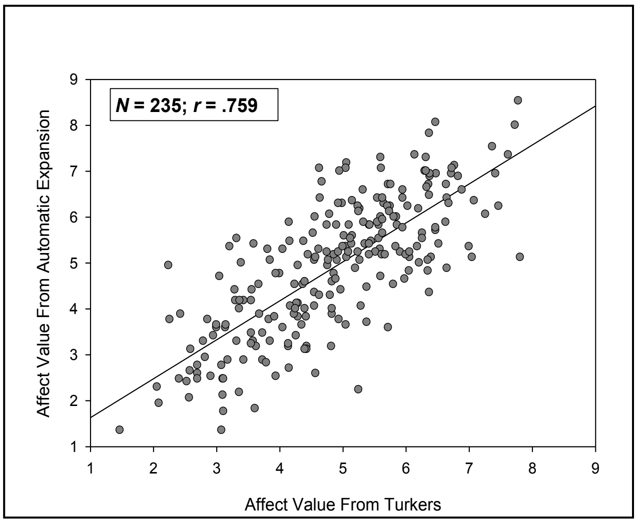

Figure 4. Scatterplot of the affect values of English words from our expansion method and those obtained from mechanical turkers.

Having established that our expansion procedure is an acceptable approach to deriving affect values for single words in English, we now tested how well affect values for words in English correlate with their foreign-language translations equivalent, specifically, Spanish, Russian, and Farsi.

4. Creating Affect Lexicons in Foreign Languages

Each English word in our expanded set of words was translated into its foreign language equivalent using Google Translate API[2]. We chose Spanish, Russian and Farsi languages for the purposes of our research. In Table 1, we show the number of words in our affect lexicons created using this procedure. The total number of words derived in the foreign languages varies, due to the fact that the Google Translate API is unable to provide a translation for a small proportion of words. The foreign language translation of a given word was assigned the same affect score as the original English word. In cases where multiple English words were translated to the same foreign language word, the average score was assigned, using an equivalent procedure to that described in Figure 2 above.

	Number of words in synonym expansion	Number of words in all synonym+ hyponym expansion
English	22,756	109,752
Spanish	17,273	107,143
Russian	17,455	107,217
Farsi	17,050	106,585

Table 1. Number of words in affect lexicons using our expansion method and translation procedure for foreign languages.

Before establishing the validity of affect scores of translated words, we wanted to determine whether the translations provided by Google Translate API were, in fact, accurate. For Spanish, a lexicon of affective scores for 1,304 words already exists (Redondo et al., 2007), and could be used for compare Google Translate output. We were able to match 88% (906/1,304) of the words. The 12% error rate was due to errors in differences in part of speech (e.g., Google Translate provided the verb form of the word, whereas Redondo et al. used the noun form). To test the accuracy of Russian and Farsi automatic translations, we selected 240 words and had trained native speakers of each language verify the translation provided by Google Translate. The percentage of English words that was incorrectly translated to Russian and Farsi were quite small, 5.4% and 10.3%, respectively.

4.1 Validation of Expanded Foreign Lexicons

Since we had a Spanish human gold standard in the Redondo lexicon to compare against, we were able to validate our automatically generated Spanish lexicon in the same manner as English expansion. We computed the correlation of scores derived automatically by our method and those collected by human participants in the Redondo lexicon, which yielded a correlation of r = .918 (see Figure 5).

[2] https://cloud.google.com/translate/docs

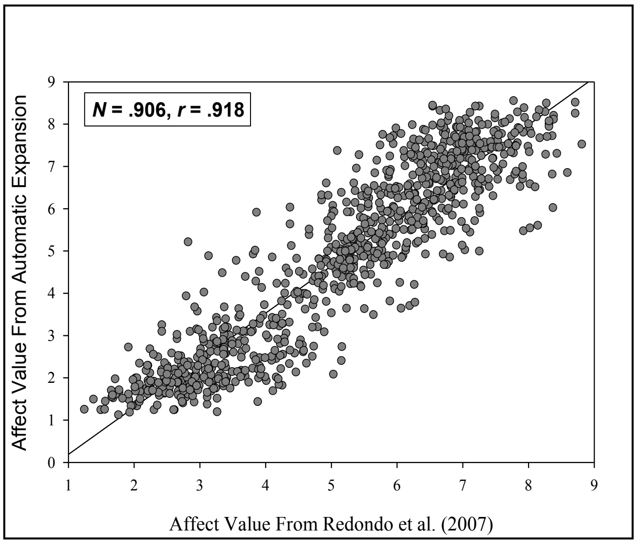

Figure 5. Scatterplot of the affect values of Spanish words from our expansion method and those reported in Redondo et al. (2007)

To provide convergent evidence of the reliability of our expansion procedure, we conducted a second validation study in which we compared the affect ratings derived using our expansion method to those given by Spanish bilinguals recruited through Amazon Mechanical Turk. (The whole task was presented to participants in Spanish.) Only Turkers who had completed at least 500 tasks with a 99% approval rate were invited to participate. Each Turker rated 240 words, one at a time, with the words being presented in a different, randomized order for each Turker. Forty words were selected from the Redondo et al. (2007) norms. The 40 words that were selected ranged from being highly negative (e.g., tóxico [toxic], terrible [terrible], entierro [burial]) to highly positive (e.g., gatito [kitten], comer [eat], miel [honey]) and served as words for which we used to determine whether the participant was fluent in Spanish and/or was taking the task seriously. That is, if a participant were to indicate ratings for these words that were highly discrepant to those reported by Redondo et al. (2007), we would exclude this participant's data from the analysis. All the words used in this validation program (as well as those used for Russian and Farsi, discussed next) were also in the second validation protocol (described above) for the English corpus. We used the same words (i.e., translation equivalents) for all languages so as to ensure that any differences in the results across the languages were not due to a different set of words used in one language but not the others. First, we considered the correlation for the ratings for the 40 words that were selected from Redondo et al.'s (2007) corpus. The correlation between the ratings given by participants in our study compared to those given by participants in Redondo et al.'s (2007) study was r = .916, p < .001. This robust correlation suggests that the ratings in our sample are reliable. We then considered the correlation between the ratings given by participants in our study compared to those derived using our automatic expansion method for the other 200 words. This analysis yielded a robust correlation of r = .851, p < .001 (see Figure 6), providing further evidence that our method of automatically computing affect ratings for Spanish words is valid.

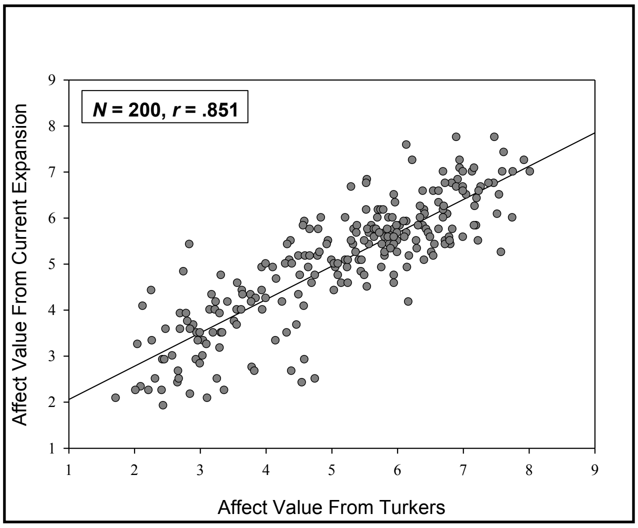

Figure 6. Scatterplot of the affect values of Spanish words from our method and those obtained from mechanical turkers.

There is no existing resource for Russian and Farsi affective norms. As a result, we ran only the validation using workers recruited through Amazon Mechanical Turk. Fourteen Turkers who are fluent in Russian, and 5 Turkers who are fluent in Farsi participated in the validation experiment. Because it is more difficult to recruit Turkers who speak these two languages through Amazon Mechanical Turk, we allowed Turkers who had completed at least 100 hits with a 96% approval rate to participate in the study. To ensure that our participants were fluent in these two languages, we added a 10-item grammar test toward the end of the survey. The grammar test assessed participants' ability to detect common grammatical errors such as subject-verb agreement and word tense. For each sentence, participants had to indicate whether there was a grammatical error. (Five sentences contained an error.) Chance performance was 50%, and the data from Turkers whose score was below 60% were excluded from all analyses. Despite the small sample size (5) in our Farsi validation, when we computed the degree of agreement on the affect rating of the words among our sample, the intraclass correlation (inter-rater agreement, see McGraw & Wong, 1996; Shrout & Fleiss, 1979) yielded a coefficient of .84. (The coefficient value ranges from 0-1, with a higher value indicating greater agreement. A value of .70 is typically accepted as good agreement; thus, our obtained value of .84 indicates that the participants showed high level of agreement in their ratings of the words.)

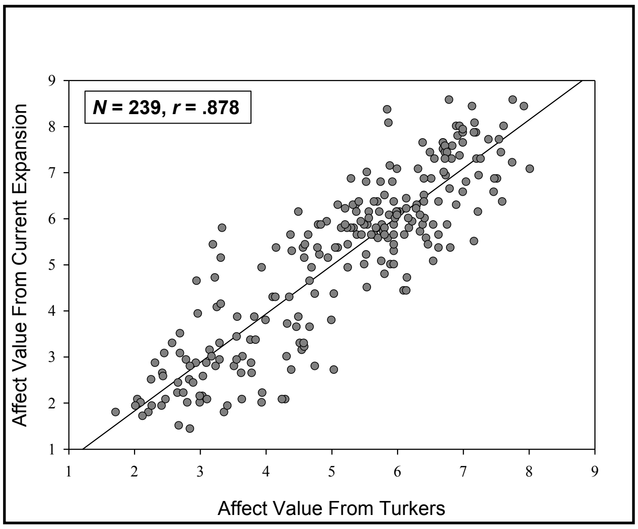

Figure 7. Scatterplot of the affect values of Russian words from our method and those obtained from mechanical turkers.

The overall correlation of the affect values given by Turkers and those derived from our expansion method was .878 for Russian (after removing one outlier; see Figure 7), and .839 for Farsi (see Figure 8). Thus these results support the conclusion that affect values for English words are translated to their Russian or Farsi equivalent, the affect values for the words are largely retained.

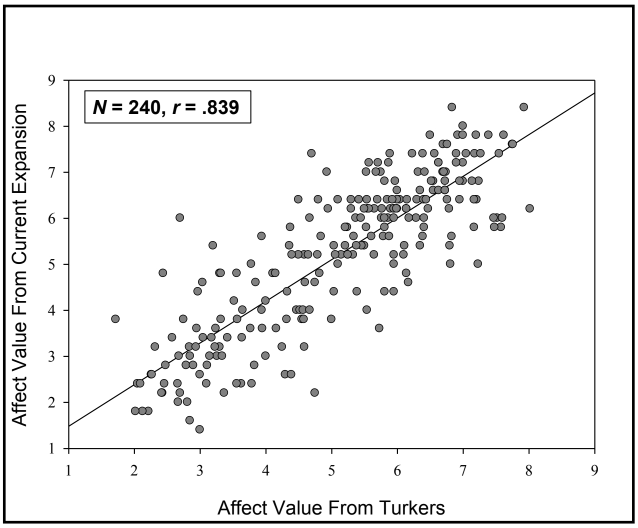

Figure 8. Scatterplot of the affect values of Farsi words from our method and those obtained from mechanical turkers.

5. Conclusion and Discussion

Overall, we obtained robust correlations in the affect ratings of words that were automatically derived compared to those obtained using human participants. In principle, the current expansion method is appropriate to use on all studies that have gathered affective norms data using human participants and methods that are both valid and reliable. Although researchers who are interested in obtaining ratings for additional words can conduct their own normative study using procedures similar to those employed by Bradley and Lang (2009) and Warriner et al. (2013), which had a group of participants rate each word on its affect, such a procedure is not ideal because it may require a lot of resources. For example, Warriner et al.'s (2013) normative study of 14,000 words collected data from as many as 1,827 participants. Thus, valid and reliable methods to automatically compute affect ratings, an approach that we used to create the present corpus, is clearly a more desirable option because it requires fewer resources.

Our results also showed that the results from our method of expansion are generalizable to words in Spanish, Russian, and Farsi. At present, there is a very small corpus (about 1,000 words; see Redondo et al. (2007) for affect values for Spanish words, and there are no resources for Russian and Farsi words. Thus, the results of the present study should be of high interest to the scientific community. However, it should be noted that because we used Google Translate (rather than ex-pert linguists) to translate English words to their foreign-language equivalent, we do not anticipate that all words will be translated accurately and thus the affect values for these words may be inaccurate. Based on the results of our study, we estimate that no more than 10% of the words will be incorrectly translated.

Our expansion technique also raises interesting questions for future researchers to investigate. One question to consider is whether our expansion method is also valid for other psychological constructs that have been collected using human participants. For example, the dimension of arousal (i.e., the intensity of emotion evoked by a word) is one variable that is of interest to many researchers. Another question is whether compound words (e.g., “holy scripture”) or short phrases (e.g., “word of god”) derived from our expansion method correlate with their source words (i.e., “bible”).

6. Bibliographical References

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. In *Proceedings of LREC '08, 6th Language Resources and Evaluation Conference, 496-500,* ELRA, Marrakech, Morocco.

Bestgen, Y. & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods.* doi: 10.3758/s13428-012-0195-z

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

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC'06, 5th Language Resources and Evaluation Conference (pp. 417-422).*

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of COLING '04, the 20th International Conference on Computational Linguistics,* 1367-1373. Geneva, Switzerland. doi:10.3115/1220355.1220555

Liu, Ting, Kit Cho, George Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz, Ignacio Cases, Ching-sheng Lin. (2014) Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*

Miller, G. A. (1995). WordNet: A Lexical database for English. Communications of the ACM, 38(11): 39-41.

Nielsen, Finn Arup. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Arxiv preprint arXiv:1103.2903.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior Research Methods, 39*, 600-605.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). *Lexicon-based methods for sentiment analysis.* Computational Linguistics. 37(2): p. 267-307.

Turney, P. D., & Littman, M. (2002). *Unsupervised learning of semantic orientation from a hundred-billion-word corpus.* Technical Report ERB-1094 (NRC-44929). National Research Council Canada.

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