# Word Affect Intensities

**Saif M. Mohammad**
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

## Abstract

Words often convey affect—emotions, feelings, and attitudes. Lexicons of word–affect association have applications in automatic emotion analysis and natural language generation. However, existing lexicons indicate only coarse categories of affect association. Here, for the first time, we create an affect intensity lexicon with real-valued scores of association. We use a technique called best–worst scaling that improves annotation consistency and obtains reliable fine-grained scores. The lexicon includes terms from general English as well as terms specific to social media communications. It has close to 6,000 entries for four basic emotions. We will be adding entries for other affect dimensions shortly.

## 1 Introduction

Words often convey affect—emotions, feelings, and attitudes. Some words have affect as a core part of their meaning. For example, *dejected* and *wistful* denotate some amount of sadness (and are thus associated with sadness). On the other hand, some words are associated with affect even though they do not denotate affect. For example, *failure* and *death* describe concepts that are usually accompanied by sadness and thus they connotate some amount of sadness. Lexicons of word–affect association have numerous applications, including: tracking brand and product perception, tracking support for issues and policies, tracking public health and well-being, literary analysis, developing more natural dialogue systems, and disaster/crisis management. However, existing manually created affect lexicons only indicate coarse categories of emotion association, for example, associated with fear or not associated with fear.

On the other hand, words can be associated with different intensities (or degrees) of an emotion. For example, most people will agree that the word condemn is associated with a greater degree of anger (or more anger) than the word irritate. However, annotating instances for fine-grained degrees of affect is a substantially more difficult undertaking than categorical annotation. Respondents are presented with greater cognitive load and it is particularly hard to ensure consistency (both across responses by different annotators and within the responses produced by the same annotator).

*Best-Worst Scaling (BWS)* is an annotation scheme that addresses these limitations (Louviere, 1991; Cohen, 2003; Louviere et al., 2015). Annotators are given $n$ items (an $n$-tuple, where $n > 1$ and commonly $n = 4$). They are asked which item is the *best* (highest in terms of the property of interest) and which is the *worst* (least in terms of the property of interest). When working on 4-tuples, best–worst annotations are particularly efficient because each best and worst annotation will reveal the order of five of the six item pairs. For example, for a 4-tuple with items A, B, C, and D, if A is the best, and D is the worst, then A $>$ B, A $>$ C, A $>$ D, B $>$ D, and C $>$ D.

We can calculate real-valued scores of association between the items and the property of interest from the best–worst annotations for a set of 4-tuples (Orme, 2009; Flynn and Marley, 2014). It has been empirically shown that three annotations each for $2N$ 4-tuples is sufficient for obtaining reliable scores (where N is the number of items) (Louviere, 1991; Kiritchenko and Mohammad, 2016).[1]

[1] At its limit, when $n = 2$, BWS becomes a *paired comparison* (Thurstone, 1927; David, 1963), but then a much larger set of tuples need to be annotated (closer to $N^2$).

Here, for the first time, we create an affect intensity lexicon with real-valued scores of association using best–worst scaling. For a given word and emotion X, the scores range from 0 to 1. A score of 1 means that the word conveys the highest amount of emotion X. A score of 0 means that the word conveys the lowest amount of emotion X. We will refer to this lexicon as the *NRC Affect Intensity Lexicon*. It has close to 6,000 entries for four basic emotions: anger, fear, joy, and sadness. We will shortly be adding entries for four more emotions: trust, disgust, anticipation, and surprise. We will also be adding entries for valence, arousal, and dominance. It includes common English terms as well as terms that are more prominent in social media platforms, such as Twitter. It includes terms that are associated with emotions to various degrees. For a given emotion, this even includes some terms that may not predominantly convey that emotion (or that convey an antonymous emotion), and yet tend to co-occur with terms that do. Antonymous terms tend to co-occur with each other more often than chance, and are particularly problematic when one uses automatic co-occurrence-based statistical methods to capture word–emotion connotations. Thus, it is particularly beneficial to have manual annotations of affect intensity for these terms.

We show that repeat annotations of the terms in the Affect Intensity Lexicon with independent annotators lead to affect association scores that are close to the scores obtained originally (Spearman Rank correlations of 0.92; Pearson correlation: 0.91). The fine-grained score obtained with BWS and the high correlations on repeat annotations indicate that BWS is both markedly discriminative (helps identify small differences in affect intensity) and markedly reliable (provides stable outcomes). We make the NRC Affect Intensity Lexicon freely available for, non-commercial, research purposes.[2]

## 2 Related Work

Psychologists have argued that some emotions are more basic than others (Ekman, 1992; Plutchik, 1980; Parrot, 2001; Frijda, 1988).[3] Thus, most work on capturing word–emotion associations has focused on a handful of emotions, especially since manually annotating for a large number of emotions is arduous. In this project, the goal is to create an affect intensity lexicon for the eight emotions emotions: anger, fear, joy, sadness, disgust, trust, anticipation, and surprise. These are the eight emotions considered to be most basic by (Plutchik, 1980). The eight emotions include the six emotions considered most basic by (Ekman, 1992), as well as trust and anticipation.

There is a large body of work on creating valence or sentiment lexicons, including the General Inquirer (Stone et al., 1966), ANEW (Nielsen, 2011; Bradley and Lang, 1999), MPQA (Wiebe et al., 2005), and norms lexicon by Warriner et al. (2013). The work on creating lexicons for categorical emotions such as joy, sadness, fear, etc, is comparatively small. WordNet Affect Lexicon (Strapparava and Valitutti, 2004) has a few hundred words annotated with the emotions they evoke.[4] It was created by manually identifying the emotions of a few seed words and then marking all their WordNet synonyms as having the same emotion. The NRC Emotion Lexicon was created by crowdsourcing and it includes entries for about 14,000 words and eight Plutchik emotions (Mohammad and Turney, 2013, 2010).[5] It also includes entries for positive and negative sentiment.

All of the emotion work and a vast majority of the valence (sentiment) work has used categorical annotation or a coarse rating scale to obtain annotations. This is not surprising, because it is difficult for humans to provide direct scores at a fine granularity. A common problem is inconsistencies in annotations among different annotators. One annotator might assign a score of 7.9 to a word, whereas another annotator may assign a score of 6.2 to the same word. It is also common that the same annotator assigns different scores to the same word at different points in time. Further, annotators often have a bias towards different parts of the scale, known as *scale region bias*. Despite this, a key question is whether humans are able to distinguish affect at only four or five coarse levels, or whether we can discriminate across much smaller affect intensity differences.

*Best-Worst Scaling (BWS)* was developed by Louviere (1991), building on some groundbreaking research in the 1960s in mathematical psychology and psychophysics by Anthony A. J.

[2] www.saifmohammad.com/WebPages/AffectIntensity.htm

[3] However, they disagree on which emotions (and how many) should be classified as basic emotions—some propose 6, some 8, some 20, and so on.

[4] http://wndomains.fbk.eu/wnaffect.html

[5] http://www.purl.org/net/saif.mohammad/research

Marley and Duncan Luce. However, it is not well known outside the areas of choice modeling and marketing research. Within the NLP community, BWS has thus far been used for creating datasets for relational similarity (Jurgens et al., 2012), word-sense disambiguation (Jurgens, 2013), and word–sentiment intensity (Kiritchenko and Mohammad, 2016). In this work we use BWS to annotate words for intensity (or degree) of affect. With BWS we address the challenges of direct scoring, and produce more reliable emotion intensity scores. Further, this will be the first dataset that will also include emotion scores for words common in social media.

There is growing work on automatically determining word–emotion associations (Mohammad and Kiritchenko, 2015; Mohammad, 2012; Strapparava and Valitutti, 2004; Yang et al., 2007). These automatic methods often assign a real-valued score representing the degree of association. However, they have been evaluated on the class of emotion they assign to each word. With the NRC Affect Intensity Lexicon, one can evaluate how accurately the automatic methods capture affect intensity.

## 3 NRC Affect Intensity Lexicon

We now describe how we created the NRC Affect Intensity Lexicon.

### 3.1 Term Selection

We chose to annotate commonly used English terms, as well as terms common in social media texts, so that the resulting lexicon can be applied widely. Twitter has a large and diverse user base, which entails rich textual content.[6] Tweets have plenty of non-standard language such as emoticons, emojis, creatively spelled words (*happee*), hashtags (*#takingastand, #lonely*) and conjoined words (*loveumom*). Tweets are often used to convey one's emotions, opinions towards products, and stance over issues. Thus, emotion analysis of tweets is particularly compelling.

Since most words do not convey a particular emotion to a marked degree, annotating all words for all emotions is sub-optimal. Thus, for each of the eight emotions, we created separate lists of terms that satisfied either one of the two properties listed below:

- The word is already known to be associated with the emotion (although the intensity of emotion it conveys is unknown).
- The word has a tendency to occur in tweets that express the emotion.

With these properties in mind, for our annotation, we included terms form two separate sources:

- The words listed in the NRC Emotion Lexicon that are marked as being associated with any of the Plutchik emotions.
- The words that tend to co-occur more often than chance with emotion-word hashtags in a large tweets corpus. (Emotion-word hashtags, such as *#angry, #fear,* and *#happiness*, act as noisy labels of the corresponding emotions.)

Since the NRC Emotion Lexicon (Mohammad and Turney, 2013, 2010) included only those terms that occur frequently in the Google n-gram corpus (Brants and Franz, 2006), these terms satisfy the 'commonly used terms' criterion as well.

As the Twitter source, we make use of the Hashtag Emotion Corpus (Mohammad, 2012), which is a large collection of tweets that each have at least one emotion-word hashtag. This dataset has emotion word hashtags corresponding to the eight basic Plutchik emotions. As mentioned before, we consider the emotion word hashtags as (noisy) labels of the corresponding emotions. For every word $w$ that occurred more than ten times in the corpus, we compute the pointwise mutual information (PMI) between the word and each of the emotion labels $e$.

$$PMI(w, e) = \log \frac{freq(w, e)}{freq(w) * freq(e)} \qquad (1)$$

where $freq(w, e)$ is the number of times $w$ occurs in a sentence with label $e$. $freq(w)$ and $freq(e)$ are the frequencies of $w$ and $e$ in the corpus. If a word has a greater-than-chance tendency to occur in tweets with a particular emotion label, then it will have a PMI score that is greater than 1. For each emotion, we included all terms in the Hashtag Emotion Corpus (Mohammad, 2012) that had a PMI $> 1$. Note that this set of terms included both terms that are more common in social media communication (for example, *soannoyed, grrrrr, stfu*, and *thx*) as well as regular English words.[7]

[6] Twitter is an online social networking and microblogging service where users post and read messages that are up to 140 characters long. The posts are called tweets.

[7] Some of the terms included from the Twitter source were deliberate spelling variations of English words, for example, *bluddy* and *sux*.

### 3.2 Annotating for Affect Intensity with Best–Worst Scaling

For each emotion, the annotators were presented with four words at a time (4-tuples) and asked to select the word that conveys the highest emotion intensity and the word that conveys the lowest emotion intensity. $2 \times N$ (where $N$ is the number of words to be annotated) distinct 4-tuples were randomly generated in such a manner that each word is seen in eight different 4-tuples, and no two 4-tuples have more than two items in common. We used the script provided by Kiritchenko and Mohammad (2016) to obtain the BWS annotations.[8]

Kiritchenko and Mohammad (2016) showed that using just three annotations per 4-tuple produces highly reliable results. We obtained four independent annotations for each 4-tuple. Note that since each word occurs in eight different 4-tuples, each word is involved in $8 \times 4 = 32$ best–worst judgments. We obtained annotations from native speakers of English residing in the United States of America. Annotators were free to provide responses to as many 4-tuples as they wished. The set of 4-tuples for each emotion was annotated by 50 to 75 people. A sample questionnaire is shown below.

---

**Words Associated With Most And Least Anger**

Words can be associated with different degrees of an emotion. For example, most people will agree that the word condemn is associated with a greater degree of anger (or more anger) than the word irritate. The goal of this task is to determine the degrees of anger associated with words. Since it is hard to give a numerical score indicating the degree of anger, we will give you four different words and ask you to indicate to us:

- which of the four words is associated with the MOST anger
- which of the four words is associated with the LEAST anger

A rule of thumb that may be helpful is that a word associated with more anger tends to occur in many angry sentences, whereas a word associated with less anger tends to occur in fewer angry sentences.

Content Warning: Since this task is about words associated with anger, some of the words you may encounter may be offensive.

**Important Notes**

- Choose the answer that you think most native speakers of English will choose.
- If the answer could be either one of two or more words (i.e., they are associated with equal degrees of anger), then select any one of them as the answer.
- Some words such as, furious and irritated, are not only associated with anger, they also explicitly express anger. Others do not express anger, but they are associated with the emotion; for example, argument and corruption are associated with anger. To be selected as 'associated with MOST anger' or 'associated with LEAST anger', a word does not have to explicitly express anger.
- Some words have more than one meaning, and the different meanings may be associated with different degrees of anger. If one of the meanings of the word is strongly associated with anger, then base your response on that meaning of the word. For example, if one of the words in the list is mad, then base your response on the angry sense of mad, as opposed to the mentally unstable sense of mad.
- Even when considering a particular sense or meaning of word, the word may convey differing degrees of anger in differing contexts. Base your response on the average anger associated with the word in that sense.
- Most importantly, try not to over-think the answer. Let your instinct guide you.

**EXAMPLE**

Q1. Identify the term associated with the MOST anger.

- tree
- grrr
- boiling
- vexed

Ans: boiling

Q2. Identify the term associated with the LEAST anger

- tree
- grrr
- boiling
- vexed

Ans: tree

---

The questionnaires for other emotions are similar in structure.

[8] http://saifmohammad.com/WebPages/BestWorst.html

| Word | Anger | Word | Fear | Word | Joy | Word | Sadness |
|---|---|---|---|---|---|---|---|
| *outraged* | 0.964 | *horror* | 0.923 | *sohappy* | 0.868 | *sad* | 0.844 |
| *brutality* | 0.959 | *horrified* | 0.922 | *superb* | 0.864 | *suffering* | 0.844 |
| *satanic* | 0.828 | *hellish* | 0.828 | *cheered* | 0.773 | *guilt* | 0.750 |
| *hate* | 0.828 | *grenade* | 0.828 | *positivity* | 0.773 | *incest* | 0.750 |
| *violence* | 0.742 | *strangle* | 0.750 | *merrychristmas* | 0.712 | *accursed* | 0.697 |
| *molestation* | 0.742 | *tragedies* | 0.750 | *bestfeeling* | 0.712 | *widow* | 0.697 |
| *volatility* | 0.687 | *anguish* | 0.703 | *complement* | 0.647 | *infertility* | 0.641 |
| *eradication* | 0.685 | *grisly* | 0.703 | *affection* | 0.647 | *drown* | 0.641 |
| *cheat* | 0.630 | *cutthroat* | 0.664 | *exalted* | 0.591 | *crumbling* | 0.594 |
| *agitated* | 0.630 | *pandemic* | 0.664 | *woot* | 0.588 | *deportation* | 0.594 |
| *defiant* | 0.578 | *smuggler* | 0.625 | *money* | 0.531 | *isolated* | 0.547 |
| *coup* | 0.578 | *pestilence* | 0.625 | *rainbow* | 0.531 | *unkind* | 0.547 |
| *overbearing* | 0.547 | *convict* | 0.594 | *health* | 0.493 | *chronic* | 0.500 |
| *deceive* | 0.547 | *rot* | 0.594 | *liberty* | 0.486 | *injurious* | 0.500 |
| *unleash* | 0.515 | *turbulence* | 0.562 | *present* | 0.441 | *memorials* | 0.453 |
| *bile* | 0.515 | *grave* | 0.562 | *tender* | 0.441 | *surrender* | 0.453 |
| *suspicious* | 0.484 | *failing* | 0.531 | *warms* | 0.391 | *beggar* | 0.422 |
| *oust* | 0.484 | *stressed* | 0.531 | *gesture* | 0.387 | *difficulties* | 0.421 |
| *ultimatum* | 0.439 | *disgusting* | 0.484 | *healing* | 0.328 | *perpetrator* | 0.359 |
| *deleterious* | 0.438 | *hallucination* | 0.484 | *tribulation* | 0.328 | *hindering* | 0.359 |

Table 1: Example entries for four (of the eight) emotions in the NRC Affect Intensity Lexicon. For each emotion, the table shows every 100th and 101th entry, when ordered by decreasing emotion intensity.

The 4-tuples of words were uploaded for annotation on the crowdsourcing platform, CrowdFlower.[9] About 5% of the data was annotated internally before hand (by the author). These questions are referred to as gold questions. The gold questions are interspersed with other questions. If one gets a gold question wrong, they are immediately notified of it. If one's accuracy on the gold questions falls below 70%, they are refused further annotation, and all of their annotations are discarded. This serves as a mechanism to avoid malicious annotations. In addition, the gold questions serve as examples to guide the annotators. In a post-annotation survey, the respondents gave the task high scores for clarity of instruction (an average of 4.5 out of 5) and overall satisfaction (an average of 4.3 out of 5).

The BWS responses were translated into scores by a simple calculation (Orme, 2009; Flynn and Marley, 2014): For each item, the score is the proportion of times the item was chosen as having the most intensity minus the proportion of times the item was chosen as having the least intensity. The scores range from -1 to 1. Since degree of emotion is a unipolar scale, we linearly transform the the -1 to 1 scores to scores in the range 0 to 1. We refer to the full list of words along with their real-valued scores of affect intensity as the *NRC Affect Intensity Lexicon*. The lexicon has about 12,000 entries with about 1500 entries for each of the eight emotions. Table 1 shows some example entries from the lexicon. Figure 1 shows the histogram of word–anger intensities. Observe that the intensity scores have a normal distribution. The histograms for other emotions have a similar shape. The lexicon is made freely available for, non-commercial, research purposes.[10]

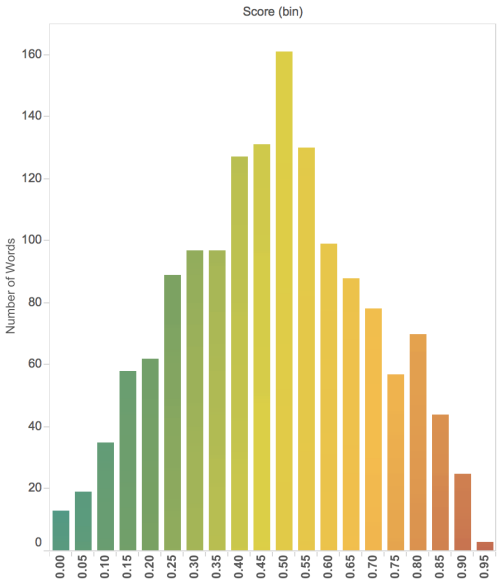


Figure 1: A histogram of word–anger intensities. Anger intensity scores are grouped in bins of size 0.05. The colors of the bars go from green to orange in increasing order of affect intensity.

[9] http://www.crowdflower.com

[10] www.saifmohammad.com/WebPages/AffectIntensity.htm

| Emotion | Spearman | Pearson |
|---|---|---|
| anger | 0.906 | 0.912 |
| fear | 0.910 | 0.912 |
| joy | 0.925 | 0.924 |
| sadness | 0.904 | 0.909 |

Table 2: Split-half reliabilities (as measured by Pearson correlation and Spearman rank correlation) for the anger, fear, joy, and sadness entries in the NRC Affect Intensity Lexicon.

## 4 Reliability of the Annotations

One cannot use standard inter-annotator agreement to determine quality of BWS annotations because the disagreement that arises when a tuple has two items that are close in emotion intensity is a useful signal for BWS. For a given 4-tuple, if respondents are not able to consistently identify the word that has highest (or lowest) emotion intensity, then the disagreement will lead to the two words obtaining scores that are close to each other, which is the desired outcome. Thus a different measure of quality of annotations must be utilized.

A useful measure of quality is reproducibility of the end result—if repeated independent manual annotations from multiple respondents result in similar intensity scores, then one can be confident that the scores capture the true emotion intensities. To assess this reproducibility, we calculate average *split-half reliability (SHR)* over 100 trials. SHR is a commonly used approach to determine consistency in psychological studies, that we employ as follows. All annotations for an item (in our case, tuples) are randomly split into two halves. Two sets of scores are produced independently from the two halves. Then the correlation between the two sets of scores is calculated. If the annotations are of good quality, then the correlation between the two halves will be high. Table 2 shows the split-half reliabilities for the anger, fear, joy, and sadness entries in the NRC Affect Intensity Lexicon. Observe that both the Pearson correlation and the Spearman rank correlations are above 0.9, indicating a high degree of reproducibility. Note that SHR indicates the quality of annotations obtained when using only half the number of annotations, the correlations obtained when repeating the experiment with four annotations for each 4-tuple is expected to be higher than 0.91. Thus 0.91 is a lower bound on the quality of annotations obtained with four annotations per 4-tuple.

## 5 Applications

The NRC Affect Intensity Lexicon has many applications including automatic sentiment and emotion analysis. Mohammad and Bravo-Marquez (2017) show its usefulness for automatically determining in intensity of emotion conveyed by tweets. They also annotate a dataset of tweets for degree (or intensity) of emotion felt by the speaker—the *Tweet Emotion Intensity Dataset*. The lexicon along with Tweet Emotion Intensity Dataset can be used to study the interplay between tweet emotion intensity and the intensity of words that make up the tweet. The lexicon also has applications in the areas of digital humanities and literary analysis, where it can be used to identify high-intensity words. The NRC Affect Intensity Lexicon can also be used as a source of gold intensity scores to evaluate automatic methods of determining word affect intensity.

## 6 Conclusions

We created the *NRC Affect Intensity Lexicon*, which is a high-coverage lexicons that captures word–affect intensities for eight basic emotions. We used the technique of best–worst scaling (BWS) to obtain fine-grained scores (and word rankings) and address issues of annotation consistency that plague traditional rating scale methods of annotation. We show that repeat annotations of the terms in the Affect Intensity Lexicon with independent annotators lead to affect association scores that are close to the scores obtained originally (Spearman Rank correlations of 0.92; Pearson correlation: 0.91). The fine-grained score obtained with BWS and the high correlations on repeat annotations indicate that BWS is both markedly discriminative (helps identify small differences in affect intensity) and markedly reliable (provides stable outcomes). The lexicon has applications in automatic emotion analysis as well as in understanding affect composition—how affect of a sentence is impacted by the affect of its constituent words. We are already in the process of adding entries for the emotions of disgust, trust, surprise, and anticipation. Then we will obtain valence, arousal, and dominance scores for all of the terms in the NRC Affect Intensity lexicon.

## Acknowledgments

Many thanks to Svetlana Kiritchenko and Tara Small for helpful discussions.

## References


Margaret M Bradley and Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. *Linguistic Data Consortium* .

Steven H. Cohen. 2003. Maximum difference scaling: Improved measures of importance and preference for segmentation. Sawtooth Software, Inc.

Herbert Aron David. 1963. *The method of paired comparisons*. Hafner Publishing Company, New York.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6(3):169–200.

T. N. Flynn and A. A. J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, Edward Elgar Publishing, pages 178–201.

Nico H Frijda. 1988. The laws of emotion. *American psychologist* 43(5):349.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Atlanta, GA, USA.

David Jurgens, Saif M. Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation*. Montréal, Canada, pages 356–364.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. San Diego, California.

Jordan J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Saif Mohammad. 2012. #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*. Montréal, Canada, pages 246–255.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Submitted*.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31(2):301–326. https://doi.org/10.1111/coin.12024.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. LA, California.

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

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC Workshop on 'Making Sense of Microposts': Big things come in small packages*. Heraklion, Crete, pages 93–98.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. Sawtooth Software, Inc.

W Parrot. 2001. *Emotions in Social Psychology*. Psychology Press.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1(3):3–33.

Philip Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, Portugal, pages 1083–1086.

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological review* 34(4):273.

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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2-3):165–210.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. pages 133–136.