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. Further, different words can convey affect to various degrees (intensities). However, existing manually created lexicons for basic emotions (such as anger and fear) indicate only coarse categories of affect association (for example, associated with anger or not associated with anger). Automatic lexicons of affect provide fine degrees of association, but they tend not to be accurate as human-created lexicons. Here, for the first time, we present a manually created affect intensity lexicon with real-valued scores of intensity for four basic emotions: anger, fear, joy, and sadness. (We will subsequently add entries for more emotions such as disgust, anticipation, trust, and surprise.) We refer to this dataset as the *NRC Affect Intensity Lexicon*, or *AIL* for short. AIL has entries for close to 6,000 English words. We used a technique called best–worst scaling (BWS) to create the lexicon. BWS improves annotation consistency and obtains reliable fine-grained scores (split-half reliability > 0.91). We also compare the entries in AIL with the entries in the *NRC VAD Lexicon*, which has valence, arousal, and dominance (VAD) scores for 20K English words. We find that anger, fear, and sadness words, on average, have very similar VAD scores. However, sadness words tend to have slightly lower dominance scores than fear and anger words. The Affect Intensity Lexicon has applications in automatic emotion analysis in a number of domains such as commerce, education, intelligence, and public health. AIL is also useful in the building of natural language generation systems.

Keywords: emotion intensity, emotion lexicon, emotion analysis, crowdsourcing, best–worst scaling, sentiment analysis

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, and developing more natural dialogue systems. Past work on manually compiling affect lexicons has focused on denotative words (Wiebe et al., 2005; Francisco and Gervás, 2006; Strapparava and Valitutti, 2004). A notable exception to this is the NRC Emotion Lexicon, which includes words that are associated with (or connotate) an emotion (Mohammad and Turney, 2013).

Words can be associated with different intensities (or degrees) of an emotion. For example, most people will agree that the word *outrage* is associated with a greater degree of anger (or more anger) than the word *irritate*. However, existing manually created affect lexicons for basic emotions such as anger and fear do not provide scores for the intensity of the emotion. Annotating instances for fine-grained intensity of affect is a substantially more difficult undertaking than categorical annotation. 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 by employing comparative annotations (Louviere, 1991; Cohen, 2003; Louviere et al., 2015; Kiritchenko and Mohammad, 2017). Annotators are given n items at a time (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 (i.e., 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). The scores can be used to rank items by the degree of association with the property of interest. 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] Kiritchenko and Mohammad (2017) showed through empirical experiments that BWS produces more reliable and more discriminating scores than those obtained using rating scales.

Here, for the first time, we create an affect intensity lexicon with real-valued scores of association for four basic emotions (anger, fear, joy, and sadness) 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 intensity (amount) of emotion X. A score of 0 means that the word conveys the lowest intensity (amount) of emotion X. We will refer to this lexicon as the *NRC Affect Intensity Lexicon (AIL)*. AIL includes entries for close to 6,000 English words. 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 associ-

[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).

ated 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 scores 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 also compare the entries in AIL with the entries in the *NRC VAD Lexicon*, which has valence, arousal, and dominance (VAD) scores for 20K English words. We find that anger, fear, and sadness words, on average, have very similar VAD scores. However, sadness words tend to have slightly lower dominance scores than fear and anger words. The Affect Intensity Lexicon has applications in automatic emotion analysis in a number of domains such as commerce, education, intelligence, and public health. AIL is also useful in the building of natural language generation systems. We have made the NRC Affect Intensity Lexicon freely available for, non-commercial, research purposes.[2]

We begin with a brief overview of the related work (Section 2), followed by a description of how we created the NRC Affect Intensity lexicon (Section 3). In Section 4, we study the valence, arousal, and dominance scores of words in the Affect Intensity Lexicon. In Section 5, we present experiments on the reliability of the annotations. In Section 6, we outline various applications of the NRC Affect Intensity lexicon. Finally, in Section 6, we present concluding remarks.

2. Related Work

Psychologists have argued that some emotions are more basic than others (Ekman, 1992; Plutchik, 1980; Frijda, 1988; Parrot, 2001).[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, we focus on four emotions common among the many proposals for basic emotions (Plutchik, 1980; Ekman, 1992; Parrot, 2001): anger, fear, joy, and sadness.

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), NRC VAD Lexicon by (Mohammad, 2018), and the 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; Mohammad and Turney, 2010).[5] It also includes entries for positive and negative sentiment.

Most prior work in sentiment analysis describes machine learning systems trained and tested on data with coarse categorical annotations. This is not surprising, because it is difficult for humans to directly provide valence (sentiment) 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 ground-breaking research in the 1960's in mathematical psychology and psychophysics by Anthony A. J. 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) and word-sense disambiguation (Jurgens, 2013). Mohammad (2018) used best–worst scaling to annotate about 20K words for valence, arousal, and dominance. In this work, we use BWS to annotate words for intensity (or degree) of basic emotions. 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). These automatic methods often assign a real-valued score representing the degree of association. Further, these association scores are likely to be somewhat correlated with the intensity of the emotion. The Affect Intensity Lexicon can be used to judge the quality of the automatic lexicons, and also to explore the extent of correlation between emotion association and emotion intensity.

[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

3. NRC Affect Intensity Lexicon

We now present how we created the NRC Affect Intensity Lexicon. The two sub-sections below describe how we chose the terms to be annotated and how we annotated the chosen terms, respectively.

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. Therefore, apart from common English terms, we also chose to annotate terms common in tweets.

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 from 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; Mohammad and Turney, 2010) includes 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.

The Hashtag Emotion Corpus (Mohammad, 2012) has tweets that each have at least one emotion-word hashtag. The 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 that occurred more than ten times in the corpus, we computed the pointwise mutual information (PMI) between the word and each of the emotion labels. 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 0. 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]

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 had more than two items in common. We used the script provided by Kiritchenko and Mohammad (2016) to obtain the 4-tuples to be annotated.[8] 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:

- the word that is associated with the MOST anger
- the word that 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.

Important Notes

- 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.

---

[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 tweets were deliberate spelling variations of English words, for example, *bluddy* and *sux*.

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

Dataset	#words	Location of Annotators	Annotation Item	#Items	#Annotators	MAI	#Q/Item	#Best–Worst Annotations
anger	1,483	USA	4-tuple of words	2,966	119	4	2	12,212
fear	1,765	USA	4-tuple of words	3,530	82	4	2	14,129
joy	1,268	USA	4-tuple of words	2,536	76	4	2	10,365
sadness	1,298	USA	4-tuple of words	2,596	76	4	2	10,429
Total	**5,814**							**47,135**

Table 1: Summary details of the current annotations done for the NRC Affect Intensity Lexicon. MAI = median number of annotations per item. Q = questions.

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.

We setup four crowdsourcing tasks corresponding to the four basic emotions. The 4-tuples of words were uploaded for annotation on the crowdsourcing platform, CrowdFlower.[9] 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 annotation tasks were approved by the National Research Council Canada's Institutional Review Board, which reviewed the proposed methods to ensure that they were ethical.

About 5% of the data was annotated internally beforehand (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 or random annotations. In addition, the gold questions serve as examples to guide the annotators.

Kiritchenko and Mohammad (2016) showed that using just three annotations per 4-tuple produces highly reliable results. In task settings, we specified that we needed annotations from four people for each word.[10] However, because of the way the gold questions work in CrowdFlower, they were annotated by more than four people. Nonetheless, the median number of annotations is four (same as the minimum number of annotations). A total of 47,135 pairs of responses (best and worst) were obtained (see Table 1).

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

[10] Note that since each word occurs in eight different 4-tuples, each word is involved in $8 \times 4 = 32$ best–worst judgments.

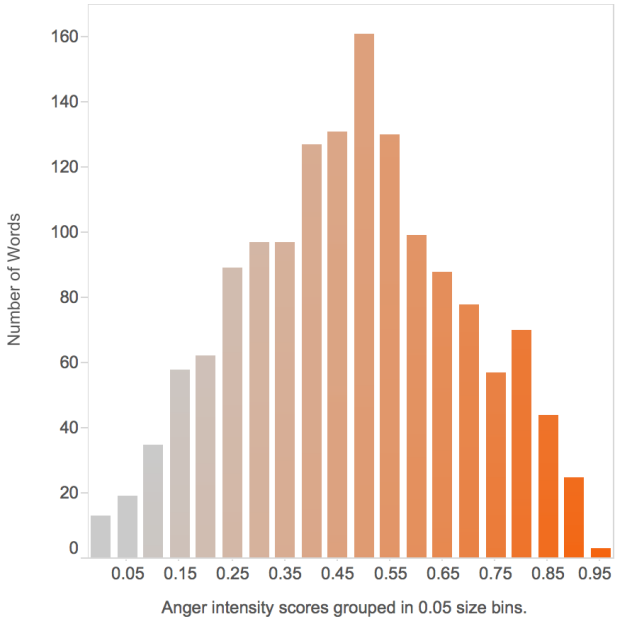

Figure 1: A histogram of word–anger intensities. Anger intensity scores are grouped in bins of size 0.05. The colors of the bars go from gray to orange in increasing order of affect intensity.

Annotation Aggregation: The intensity scores were calculated from the BWS responses using a simple counting procedure (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 (least emotion intensity) to 1 (the most emotion intensity). Since degree of emotion is a unipolar scale, we linearly transform the -1 to 1 scores to scores in the range 0 (least emotion intensity) to 1 (the most emotion intensity). We refer to the full list of words along with their real-valued scores of affect intensity as the *NRC Affect Intensity Lexicon*.

Distribution of Scores: Figure 1 shows a histogram of word–anger intensities. The words are grouped into bins of scores 0–0.05, 0.05–0.1, and so on until 0.95–1. Observe that the intensity scores have a normal distribution. The histograms for other emotions have a similar shape.

Table 1 gives a summary of the number of items annotated and the number of annotations obtained. Table 2 shows some example entries from the lexicon. The lexicon is made freely available.

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 2: Example entries for four emotions in the NRC Affect Intensity Lexicon. For each emotion, the table shows every 100th and 101st entry, when ordered by decreasing emotion intensity.

4. Relationships of the Basic Emotions with Valence, Arousal, and Dominance

Even though the basic emotions model has long enjoyed the attention of psychologists, the valence–arousal–dominance (VAD) model (Russell, 2003) is also widely accepted. According to the VAD model of affect, individual emotions are points in a three-dimensional space of valence (positiveness–negativeness), arousal (active–passive), and dominance (dominant–submissive). Both the basic emotions model and the VAD model offer different perspectives that help our understanding of emotions. However, there is little work relating the two models of emotion with each other. Much of the past work on textual utterances such as sentences and tweets, is based on exactly one or the other model (not both). For example, corpora annotated for emotions are either annotated only for the basic emotions (Strapparava and Mihalcea, 2007; Mohammad and Bravo-Marquez, 2017b) or only for valence, arousal, and dominance (Yu et al., ; Mohammad et al., 2017; Nakov et al., 2016). Mohammad and Kiritchenko (2018) created the first dataset of tweets manually annotated for multiple affect dimensions from both the basic emotion model and the VAD model. For each emotion dimension, they annotated the data for coarse classes (such as no anger, low anger, moderate anger, and high anger) and also for fine real-valued scores indicating the intensity of emotion (anger, sadness, valence, etc.). They present an analysis of emotion intensities of tweets and their relationship with valence.

Similar to the situation for textual corpora, *words* have been annotated largely either just for valence, arousal, and dominance (*ANEW* (Bradley and Lang, 1999), the *Warriner Lexicon* (Warriner et al., 2013), and the *NRC VAD Lexicon* (Mohammad, 2018)) or just for association with basic emotions (the NRC Emotion Lexicon (Mohammad and Turney, 2013; Mohammad and Turney, 2010)). Since all the words in the Affect Intensity Lexicon also have entries in the NRC VAD Lexicon (Mohammad, 2018), we now examine the relationship between the valence, arousal, and dominance scores across different basic emotions.

Emotion	Avg. Valence	Avg. Arousal	Avg. Dominance
anger	0.26	0.66	0.46
fear	0.29	0.66	0.48
joy	0.77	0.52	0.60
sadness	0.24	0.58	0.38

Figure 2: Average valence, arousal, and dominance scores for each basic emotion. The cells are in shades of green with the darkness proportional to the score: lighter shades indicate low scores and darker shades indicate high scores.

4.1. Valence, Arousal, and Dominance of Words in the Affect Intensity Lexicon

The NRC VAD Lexicon (Mohammad, 2018) has valence, arousal, and dominance scores for over 20,000 commonly used English terms. It was created using best–worst scaling in a similar approach described earlier in this paper for obtaining emotion intensity scores. The three sets of scores range from 0 (lowest valence, arousal, and dominance) to 1 (highest valence, arousal, and dominance).

For each of the words in the Affect Intensity Lexicon, we looked up their entries in the NRC VAD Lexicon for scores of valence, arousal, and dominance. Figure 2 shows the average scores for each of the basic emotions. Figure 3 shows the scatter plot of the Affect Intensity Lexicon words across the orthogonal valence-arousal space.

Observe that as expected, joy words have much higher valence scores (are much more positive) on average than the anger, fear, and sadness words. Joy words also have

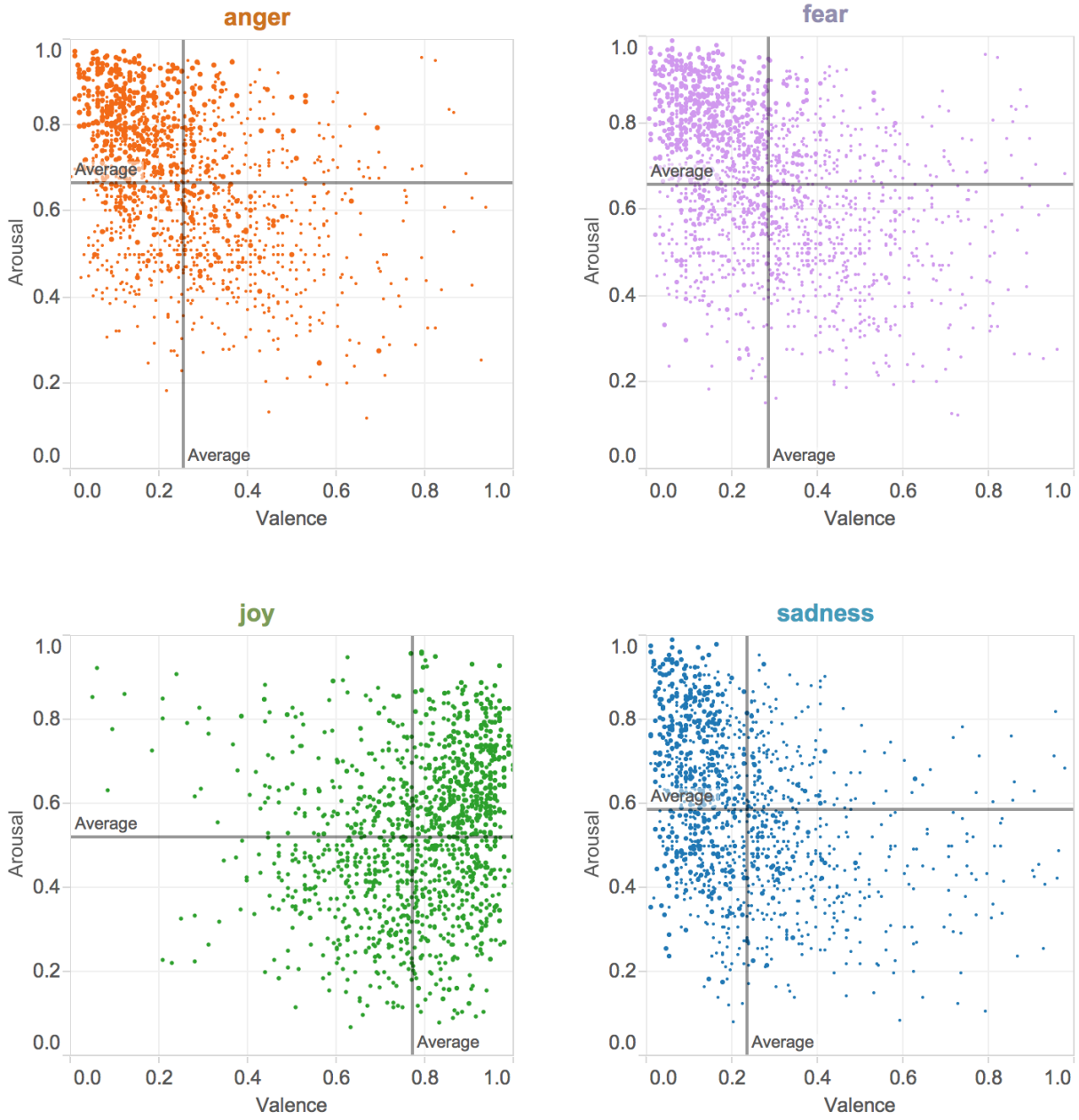

Figure 3: Valence–Arousal scatter plots for words associated with each of the four basic emotions. For example, the anger plot on the top left has points for every anger word in the Affect Intensity Lexicon. The position of the point indicates its valence and arousal scores (as obtained from the NRC VAD Lexicon). The size of the point is proportional to the intensity of anger (as obtained from AIL). The size of the point is proportional to the intensity of the corresponding emotion.

somewhat lower arousal scores (are more passive) on average than sadness words, which in turn have somewhat lower average arousal scores than anger and fear. Interestingly, anger and fear have a very similar profile of average VAD scores. Sadness words, on average, have the lowest average valence, followed by fear and anger.

To determine whether the dominance–arousal space separates the three negative emotion words from each other, we generated the corresponding scatter plots as well. See Figure 4. Observe that words conveying negative emotions can belong to a wide and overlapping range of arousal and dominance scores. The range of scores now overlaps markedly with the joy words as well. Figure 5 in the Appendix shows the scatter plots for the valence–dominance space.

Overall, we observe that the three negative emotions can be conveyed by words having a wide range of values for valence, arousal, and dominance.[11] Let the words with emotion intensity scores greater than 0.5 be called the *upper-half subset*. The upper half subset includes words expressing medium to high emotion intensity. Table 3 lists, for each emotion, the top four words that have highest and lowest valence, arousal, and dominance scores in the upper-half subset of the emotion. Note that for the negative emotions, the highest valence entries in the upper-half subset of anger are still expected to be somewhat negative.[12]

[11] The range is limited to the lower half of valence, but knowing valence is not sufficient to determine the precise basic emotion.

[12] For example, amongst the moderate-to-high anger terms, the highest valence term is still expected to be somewhat negative.

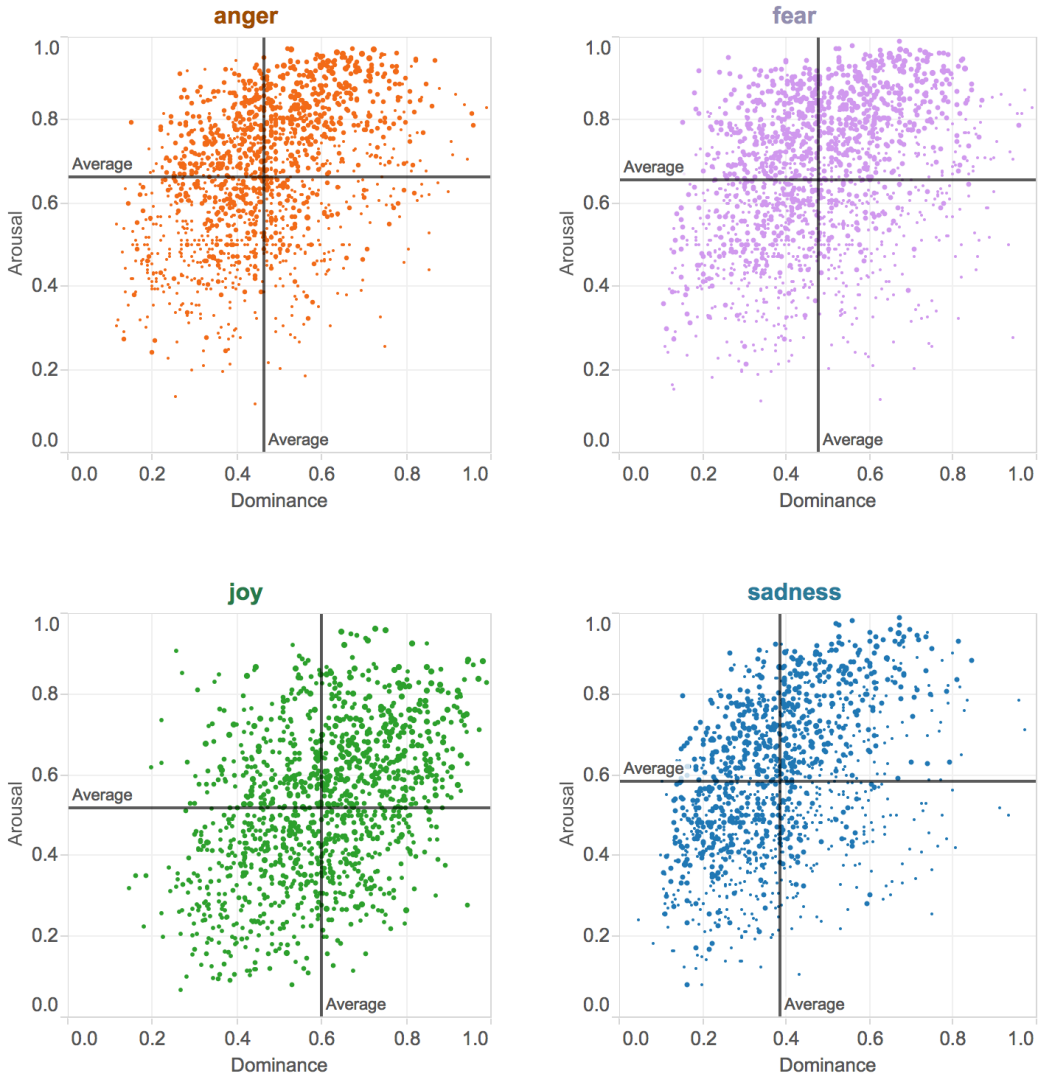

Figure 4: Dominance–Arousal scatter plots for words associated with each of the four basic emotions.

Emotion	V↑↑	V↓↓	A↑↑	A↓↓	D↑↑	D↓↓
anger > 0.5	*blaze*	*shit*	*homicide*	*batter*	*domination*	*casualty*
	glare	*homicide*	*terrorism*	*tiredofit*	*battle*	*idiots*
	incense	*murderous*	*violently*	*causality*	*overbearing*	*slave*
	temper	*terrorist*	*enraged*	*cross*	*dictatorial*	*dishonest*
fear > 0.5	*seize*	*nightmare*	*abduction*	*senile*	*domination*	*defenseless*
	meltdown	*afraid*	*exorcism*	*coma*	*projectiles*	*hopeless*
	retribution	*homicide*	*homicide*	*stalk*	*dictator*	*cowardly*
	enforce	*murderer*	*violently*	*hopeless*	*beastly*	*casualty*
joy > 0.5	*generous*	*raving*	*elated*	*stressfree*	*powerful*	*silly*
	happily	*zeal*	*excitation*	*peaceful*	*success*	*heheh*
	love	*silly*	*euphoria*	*serenity*	*triumphant*	*weeeee*
	magnificent	*boisterous*	*erotic*	*tranquility*	*winning*	*snuggles*
sadness > 0.5	*meltdown*	*bankruptcy*	*abduction*	*nothingness*	*warfare*	*defenseless*
	console	*disheartening*	*exorcism*	*alone*	*earthquake*	*weakly*
	insurmountable	*homicide*	*homicide*	*emptiness*	*bomber*	*hopeless*
	longing	*pain*	*terrorism*	*senile*	*unforgiving*	*pity*

Table 3: The top four words that have highest and lowest valence (V), arousal (A), and dominance (D) scores, while also having an emotion intensity score greater than 0.5 (in the upper-half subset). The emotion intensity scores are taken from the NRC Affect Intensity Lexicon and valence, arousal, and dominance scores are taken from the NRC VAD Lexicon. ↑↑ indicates the highest score entries. ↓↓ indicates the lowest score entries.

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

Table 4: 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.

5. 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 4 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.

6. Applications and Future Work

The NRC Affect Intensity Lexicon has many applications including automatic emotion analysis in a number of domains such as commerce, education, intelligence, and public health. The AIL was already used by several teams that participated in the WASSA-2017 shared task on Emotion Intensity in Tweet (Mohammad and Bravo-Marquez, 2017b) as well as the SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018) (including the teams that came first in both shared tasks). AIL is also useful in the building of natural language generation systems.

We are currently using the NRC Affect Intensity Lexicon along with tweets datasets that were annotated for emotion intensity (*Tweet Emotion Intensity Dataset* (Mohammad and Bravo-Marquez, 2017a)), to test the extent to which people convey strong emotions in tweets using high-intensity emotion words. We will also use the lexicon to identify syllables that consistently tend to occur in words with strong affect associations. This has implications in understanding how some syllables and sounds have a tendency to occur in words referring to semantically related concepts. Identifying emotions associated with a syllable is also useful in generating names for literary characters and commercial products.

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.

7. Conclusions

This paper describes how we created the *NRC Affect Intensity Lexicon*—a crowdsourced lexicon that captures word–affect intensities for four basic emotions: anger, fear, joy, and sadness. We used a technique called best–worst scaling (BWS) to obtain fine-grained scores (and word rankings). BWS addresses 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 (split-half reliability: $rho = 0.92$, $r = 0.91$). The fine-grained scores 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 Affect Intensity 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 will continue to add entries for other emotions such as disgust, trust, surprise, and anticipation. We will use the lexicon to study the role emotion words play in high emotion intensity tweets, using the Tweet Emotion Intensity Dataset that has intensity scores for whole tweets. We will also use the lexicon to determine syllables and phonetic sounds that are associated with particular affect categories, that is, syllables that tend to occur more often than average in affect-associated words. The lexicon is made freely available.

Acknowledgments

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

8. Bibliographical References

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

Brants, T. and Franz, A. (2006). Web 1t 5-gram version 1. *Linguistic Data Consortium*.

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

David, H. A. (1963). *The method of paired comparisons*. Hafner Publishing Company, New York.

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

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

Francisco, V. and Gervás, P. (2006). Automated mark up of affective information in English texts. In Petr Sojka, et al., editors, *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in CS*, pages 375–382. Springer.

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

Jurgens, D., Mohammad, S. M., Turney, P., and Holyoak, K. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation*, pages 356–364, Montréal, Canada, 7-8 June.

Jurgens, D. (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.

Kiritchenko, S. and Mohammad, S. M. (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.

Kiritchenko, S. and Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of The Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

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

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

Mohammad, S. M. and Bravo-Marquez, F. (2017a). Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada.

Mohammad, S. M. and Bravo-Marquez, F. (2017b). WASSA-2017 shared task on emotion intensity. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark.

Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M. and Kiritchenko, S. (2018). Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Mohammad, S. M. and Turney, P. D. (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.

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

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, USA.

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

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Submitted*.

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Stoyanov, V., and Zhu, X. (2016). Developing a successful semeval task in sentiment analysis of twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

Nielsen, F. Å. (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*, pages 93–98, Heraklion, Crete.

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

Parrot, W. (2001). *Emotions in Social Psychology*. Psychology Press.

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

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., and associates. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, Prague, Czech Republic.

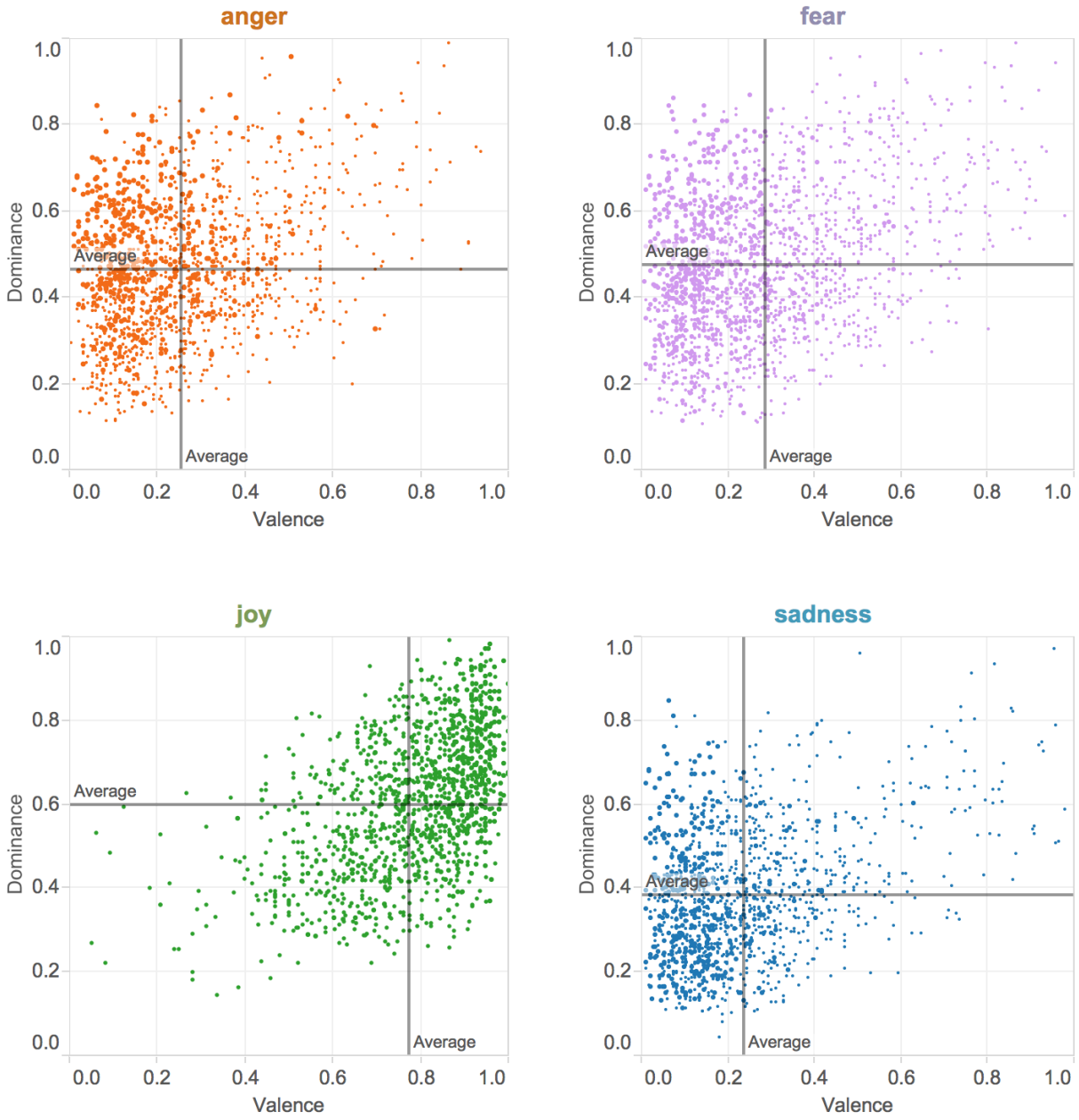

Figure 5: Valence–Dominance scatter plots for words associated with each of the four basic emotions. The size of the point is proportional to the intensity of the corresponding emotion. For example, the anger plot on the top left has points for every anger word in the Affect Intensity Lexicon. The position of the point indicates its valence and dominance scores (as obtained from the NRC VAD Lexicon). The size of the point is proportional to the intensity of anger (as obtained from the Affect Intensity lexicon).

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

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

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

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

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. R., and Zhang, X.-j.). Building Chinese affective resources in valence-arousal dimensions. 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.

9. Appendix

Figure 5 show scatter plots for the four basic emotions in the valence-dominance space.