Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words

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

Abstract

Words play a central role in language and thought. Factor analysis studies have shown that the primary dimensions of meaning are valence, arousal, and dominance (VAD). We present the NRC VAD Lexicon, which has human ratings of valence, arousal, and dominance for more than 20,000 English words. We use Best–Worst Scaling to obtain fine-grained scores and address issues of annotation consistency that plague traditional rating scale methods of annotation. We show that the ratings obtained are vastly more reliable than those in existing lexicons. We also show that there exist statistically significant differences in the shared understanding of valence, arousal, and dominance across demographic variables such as age, gender, and personality.

1 Introduction

Words are the smallest meaningful utterances in language. They play a central role in our understanding and descriptions of the world around us. Some believe that the structure of a language even affects how we think (principle of linguistic relativity aka the SapirWhorf hypothesis). Several influential factor analysis studies have shown that the three most important, largely independent, dimensions of word meaning are valence (positiveness–negativeness/pleasure–displeasure), arousal (active–passive), and dominance (dominant–submissive) (Osgood et al., 1957; Russell, 1980, 2003).[1] Thus, when comparing the meanings of two words, we can compare their degrees of valence, arousal, or dominance. For example, the word *banquet* indicates more positiveness than the word *funeral*; *nervous* indicates more arousal than *lazy*; and *fight* indicates more dominance than *delicate*.

Access to these degrees of valence, arousal, and dominance of words is beneficial for a number of applications, including those in natural language processing (e.g., automatic sentiment and emotion analysis of text), in cognitive science (e.g., for understanding how humans represent and use language), in psychology (e.g., for understanding how people view the world around them), in social sciences (e.g., for understanding relationships between people), and even in evolutionary linguistics (e.g., for understanding how language and behaviour inter-relate to give us an advantage).

Existing VAD lexicons (Bradley and Lang, 1999; Warriner et al., 2013) were created using rating scales and thus suffer from limitations associated with the method (Presser and Schuman, 1996; Baumgartner and Steenkamp, 2001). These include: inconsistencies in annotations by different annotators, inconsistencies in annotations by the same annotator, scale region bias (annotators often have a bias towards a portion of the scale), and problems associated with a fixed granularity.

In this paper, we describe how we obtained human ratings of valence, arousal, and dominance for more than 20,000 commonly used English words by crowdsourcing. Notably, we use a comparative annotation technique called *Best-Worst Scaling (BWS)* that addresses the limitations of traditional rating scales (Louviere, 1991; Cohen, 2003; Louviere et al., 2015). The scores are fine-grained real-valued numbers in the interval from 0 (lowest V, A, or D) to 1 (highest V, A, or D). We will refer to this new lexicon as the *NRC Valence, Arousal, and Dominance (VAD) Lexicon*.[2]

[1] We will refer to the three dimensions individually as *V*, *A*, and *D*, and together as *VAD*.

[2] NRC refers to National Research Council Canada.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 174–184
Melbourne, Australia, July 15 - 20, 2018.

Correlations (r) between repeated annotations, through metrics such as *split-half reliability (SHR)*, are a common way to evaluate the reliabilities of ordinal and rank annotations. We show that our annotations have SHR scores of $r = 0.95$ for valence, $r = 0.90$ for arousal, and $r = 0.91$ for dominance. These scores are well above the SHR scores obtained by Warriner et al. (2013), and indicate high reliability.

Respondents who provided valence, arousal, and dominance annotations, were given the option of additionally filling out a brief demographic questionnaire to provide details of their age, gender, and personality traits. This demographic information along with the VAD annotations allows us to determine whether attributes such as age, gender, and personality impact our understanding of the valence, arousal, and dominance of words. We show that even though overall the annotations are consistent (as seen from the high SHR scores), people aged over 35 are significantly more consistent in their annotations than people aged 35 or less. We show for the first time that men have a significantly higher shared understanding of dominance and valence of words, whereas women have a higher shared understanding of the degree of arousal of words. We find that some personality traits significantly impact a person's annotations of one or more of valence, arousal, and dominance. We hope that these and other findings described in the paper foster further research into how we use language, how we represent concepts in our minds, and how certain aspects of the world are more important to certain demographic groups leading to higher degrees of shared representations of those concepts within those groups.

All of the annotation tasks described in this paper were approved by our institution's review board, which examined the methods to ensure that they were ethical. Special attention was paid to obtaining informed consent and protecting participant anonymity. The NRC VAD Lexicon is made freely available for research and non-commercial use through our project webpage.[3]

2 Related Work

Primary Dimensions of Meaning: Osgood et al. (1957) asked human participants to rate words along dimensions of opposites such as *heavy–light, good–bad, strong–weak,* etc. Factor analysis of these judgments revealed that the three most prominent dimensions of meaning are evaluation (*good–bad*), potency (*strong–weak*), and activity (*active–passive*). Russell (1980, 2003) showed through similar analyses of emotion words that the three primary independent dimensions of emotions are valence or pleasure (positiveness–negativeness/pleasure–displeasure), arousal (active–passive), and dominance (dominant–submissive). He argues that individual emotions such as joy, anger, and fear are points in a three-dimensional space of valence, arousal, and dominance. It is worth noting that even though the names given by Osgood et al. (1957) and Russell (1980) are different, they describe similar dimensions (Bakker et al., 2014).

Existing Affect Lexicons: Bradley and Lang (1999) asked annotators to rate valence, arousal, and dominance—for more than 1,000 words—on a 9-point rating scale. The ratings from multiple annotators were averaged to obtain a score between 1 (lowest V, A, or D) to 9 (highest V, A, or D). Their lexicon, called the *Affective Norms of English Words (ANEW)*, has since been widely used across many different fields of study. More than a decade later, Warriner et al. (2013) created a similar lexicon for more than 13,000 words, using a similar annotation method. There exist a small number of VAD lexicons in non-English languages as well, such as the ones created by Moors et al. (2013) for Dutch, by Võ et al. (2009) for German, and by Redondo et al. (2007) for Spanish. The NRC VAD lexicon is the largest manually created VAD lexicon (in any language), and the only one that was created via comparative annotations (instead of rating scales).

Best-Worst Scaling: Best-Worst Scaling (BWS) was developed by (Louviere, 1991), building on work in the 1960's in mathematical psychology and psychophysics. Annotators are given n items (an n-tuple, where $n > 1$ and commonly $n = 4$).[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 (e.g., for a 4-tuple with items

[3] http://saifmohammad.com/WebPages/nrc-vad.html

[4] 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).

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). Real-valued scores of association between the items and the property of interest can be determined using simple arithmetic on the number of times an item was chosen best and number of times it was chosen worst (as described in Section 3) (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). Kiritchenko and Mohammad (2017) showed through empirical experiments that BWS produces more reliable and more discriminating scores than those obtained using rating scales. (See Kiritchenko and Mohammad (2016, 2017) for further details on BWS.)

Within the NLP community, BWS has been used for creating datasets for relational similarity (Jurgens et al., 2012), word-sense disambiguation (Jurgens, 2013), word–sentiment intensity (Kiritchenko and Mohammad, 2016), word–emotion intensity (Mohammad, 2018), and tweet–emotion intensity (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; Mohammad and Kiritchenko, 2018).

Automatically Creating Affect Lexicons: There is growing work on automatically determining word–sentiment and word–emotion associations (Yang et al., 2007; Mohammad and Kiritchenko, 2015; Yu et al., 2015; Staiano and Guerini, 2014). The VAD Lexicon can be used to evaluate how accurately the automatic methods capture valence, arousal, and dominance.

3 Obtaining Human Ratings of Valence, Arousal, and Dominance

We now describe how we selected the terms to be annotated and how we crowdsourced the annotation of the terms using best–worst scaling.

3.1 Term Selection

We chose to annotate commonly used English terms. We especially wanted to include terms that denotate or connotate emotions. We also include terms common in tweets.[5] Specifically, we include terms from the following sources:

- All terms in the NRC Emotion Lexicon (Mohammad and Turney, 2013). It has about 14,000 words with labels indicating whether they are associated with any of the eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Plutchik, 1980).
- All 4,206 terms in the positive and negative lists of the General Inquirer (Stone et al., 1966).
- All 1,061 terms listed in ANEW (Bradley and Lang, 1999).
- All 13,915 terms listed in the Warriner et al. (2013) lexicon.
- 520 words from the Roget's Thesaurus categories corresponding to the eight basic Plutchik emotions.[6]
- About 1000 high-frequency content terms, including emoticons, from the Hashtag Emotion Corpus (HEC) (Mohammad, 2012).[7]

The union of the above sets resulted in 20,007 terms that were then annotated for valence, arousal, and dominance.

3.2 Annotating VAD via Best–Worst Scaling

We describe below how we annotated words for valence. The same approach is followed for arousal and dominance. The annotators were presented with four words at a time (4-tuples) and asked to select the word with the highest valence and the word with the lowest valence. The questionnaire uses a set of paradigm words that signify the two ends of the valence dimension. The paradigm words were taken from past literature on VAD (Bradley and Lang, 1999; Osgood et al., 1957; Russell, 1980). The questions used for valence are shown below.

Q1. Which of the four words below is associated with the MOST happiness / pleasure / positiveness / satisfaction / contentedness / hopefulness OR LEAST unhappiness / annoyance / negativeness / dissatisfaction / melancholy / despair? (Four words listed as options.)

Q2. Which of the four words below is associated with the LEAST happiness / pleasure / positiveness / satisfaction / contentedness / hopefulness OR MOST unhappiness / annoyance / negativeness / dissatisfaction / melancholy / despair? (Four words listed as options.)

[5] Tweets include non-standard language such as emoticons, emojis, creatively spelled words (*happee*), hashtags (*#takingastand, #lonely*) and conjoined words (*loveumom*).

[6] http://www.gutenberg.org/ebooks/10681

[7] All tweets in the HEC include at least one of the eight basic emotion words as a hashtag word (*#anger, #sadness*, etc.).

Dataset	#words	Location of Annotators	Annotation Item	#Items	#Annotators	MAI	#Q/Item	#Best–Worst Annotations
valence	20,007	worldwide	4-tuple of words	40,014	1,020	6	2	243,295
arousal	20,007	worldwide	4-tuple of words	40,014	1,081	6	2	258,620
dominance	20,007	worldwide	4-tuple of words	40,014	965	6	2	276,170
Total								**778,085**

Table 1: A summary of the annotations for valence, arousal, and dominance. MAI = minimum number of annotations per item. Q = questions. A total of 778,085 pairs of best–worst responses were obtained.

Questions for arousal and dominance are similar.[8]

Detailed directions and example questions (with suitable responses) were provided in advance. $2 \times N$ 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 (where N is the number of words to be annotated).[9]

Crowdsourcing: We setup three separate crowdsourcing tasks corresponding to valence, arousal, and dominance. The 4-tuples of words were uploaded for annotation on the crowdsourcing platform, *CrowdFlower*.[10] We obtained annotations from native speakers of English residing around the world. Annotators were free to provide responses to as many 4-tuples as they wished. The annotation tasks were approved by our institution's review board.

About 2% of the data was annotated beforehand by the authors. These questions are referred to as gold questions. CrowdFlower interspersed the gold questions with the other questions. If a crowd worker answered a gold question incorrectly, then they were immediately notified, the annotation was discarded, and an additional annotation was requested from a different annotator. If an annotator's accuracy on the gold questions fell below 80%, then they were refused further annotation, and all of their annotations were discarded. This served as a mechanism to avoid malicious and random annotations. The gold questions also served as examples to guide the annotators.

[8]The two ends of the arousal dimension were described with the words: arousal, activeness, stimulation, frenzy, jitteriness, alertness AND unarousal, passiveness, relaxation, calmness, sluggishness, dullness, sleepiness. The two ends of the dominance dimension were described with the words: dominant, in control of the situation, powerful, influential, important, autonomous AND submissive, controlled by outside factors, weak, influenced, cared-for, guided.

[9]We used the script provided by Kiritchenko and Mohammad (2016) to generate the 4-tuples from the list of terms: http://saifmohammad.com/WebPages/BestWorst.html

[10]CrowdFlower later changed its name to Figure Eight: https://www.figure-eight.com

Dimension	Word	Score↑	Word	Score↓
valence	*love*	1.000	*toxic*	0.008
	happy	1.000	*nightmare*	0.005
	happily	1.000	*shit*	0.000
arousal	*abduction*	0.990	*mellow*	0.069
	exorcism	0.980	*siesta*	0.046
	homicide	0.973	*napping*	0.046
dominance	*powerful*	0.991	*empty*	0.081
	leadership	0.983	*frail*	0.069
	success	0.981	*weak*	0.045

Table 2: The terms with the highest (↑) and lowest (↓) valence (V), arousal (A), and dominance (D) scores in the VAD Lexicon.

In the task settings for CrowdFlower, we specified that we needed annotations from six people for each word.[11] However, because of the way the gold questions work in CrowdFlower, they were annotated by more than six people. Both the minimum and the median number of annotations per item was six. See Table 1 for summary statistics on the annotations.[12]

Annotation Aggregation: The final VAD 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 the best (highest V/A/D) minus the proportion of times the item was chosen as the worst (lowest V/A/D). The scores were linearly transformed to the interval: 0 (lowest V/A/D) to 1 (the highest V/A/D). We refer to the list of words along with their scores for valence, arousal, and dominance as the *NRC Valence, Arousal, and Dominance Lexicon*, or the *NRC VAD Lexicon* for short. Table 2 shows entries from the lexicon with the highest and lowest scores for V, A, and D.

[11]Note that since each word occurs in eight different 4-tuples, it is involved in $8 \times 6 = 48$ best–worst judgments.

[12]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).

Attribute	Value	%	Value	%
Gender	f	37	m	63
Age	≤35	70	>35	30
Personality	Ag	69	Di	31
	Co	52	Ea	48
	Ex	52	In	48
	Ne	40	Se	60
	Op	50	Cl	50

Table 3: Summary of the demographic information provided by the annotators.

4 Demographic Survey

Respondents who annotated our VAD questionnaires were given a special code through which they could then optionally respond to a separate CrowdFlower survey asking for their demographic information: age, gender, country they live in, and personality traits. For the latter, we asked how they viewed themselves across the big five (Barrick and Mount, 1991) personality traits:

- Agreeableness (Ag) – Disagreeableness (Di): friendly and compassionate or careful in whom to trust, argumentative
- Conscientiousness (Co) – Easygoing (Ea): efficient and organized (prefer planned and self-disciplined behaviour) or easy-going and carefree (prefer flexibility and spontaneity)
- Extrovert (Ex) – Introvert (In): outgoing, energetic, seek the company of others or solitary, reserved, meeting many people causes anxiety
- Neurotic (Ne) – Secure (Se): sensitive and nervous (often feel anger, anxiety, depression, and vulnerability) or secure and confident (rarely feel anger, anxiety, depression, and vulnerability)
- Open to experiences (Op) – Closed to experiences (Cl): inventive and curious (seek out new experiences) or consistent and cautious (anxious about new experiences)

The questionnaire described the two sides of the dimension using only the texts after the colons above.[13] The questionnaire did not ask for identifying information such as name or date of birth.

In total, 991 people (55% of the VAD annotators) chose to provide their demographic information. Table 3 shows the details.

	V	A	D
Ours–Warriner	0.814	0.615	0.326

Table 4: Pearson correlations between our V, A, and D scores and the Warriner scores.

Lexicon	V–A	A–D	V–D
Ours	-0.268	0.302	0.488
Ours (Warriner subset)	-0.287	0.322	0.463
Warriner	-0.185	-0.180	0.717

Table 5: Pearson correlations between various pair-wise combinations of V, A, and D.

5 Examining of the NRC VAD Lexicon

5.1 A Comparsion of the NRC VAD Lexicon and the Warriner et al. Lexicon Scores

We calculated the Pearson correlations r between the NRC VAD Lexicon scores and the Warriner et al. Lexicon scores. Table 4 shows the results. (These numbers were calculated for the 13,915 common terms across the two lexicons.) Observe that the especially low correlations for dominance and arousal indicate that our lexicon has substantially different scores and rankings of terms by these dimensions. Even for valence, a correlation of 0.81 indicates a marked amount of differences in scores.

5.2 Independence of Dimensions

Russell (1980) found through his factor analysis work that valence, arousal, and dominance are nearly independent dimensions. However, Warriner et al. (2013) report that their scores for valence and dominance have substantial correlation ($r = 0.717$). Given that the split-half reliability score for their dominance annotations is only 0.77, the high V–D correlations raises the suspicion whether annotators sufficiently understood the difference between dominance and valence. Table 5 shows the correlations between various pair-wise combinations of valence, arousal, and dominance for both our lexicon and the Warriner lexicon. Observe that unlike the Warriner annotations where V and D are highly correlated, our annotations show that V and D are only slightly correlated. The correlations for V–A and A–D are low in both our and Warriner annotations, albeit slightly higher in magnitude in our annotations.

[13] How people view themselves may be different from what they truly are. The conclusions in this paper apply to groups that view themselves to be a certain personality type.

Annotations	#Terms	#Annotations	V	A	D
a. Ours (on all terms)	20,007	6 per tuple	0.950	0.899	0.902
b. Ours (on only those terms also in Warriner)	13,915	6 per tuple	0.952	0.905	0.906
c. Warriner et al. (2013)	13,915	20 per term	0.914	0.689	0.770

Table 6: Split-half reliabilities (as measured by Pearson correlation) for valence, arousal, and dominance scores obtained from our annotations and the Warriner et al. annotations.

5.3 Reliability of the Annotations

A useful measure of quality is reproducibility of the end result—repeated independent manual annotations from multiple respondents should result in similar scores. To assess this reproducibility, we calculate average *split-half reliability (SHR)* over 100 trials. All annotations for an item (in our case, 4-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 6 shows the split-half reliabilities (SHR) for valence, arousal, and dominance annotations. Row *a.* shows the SHR on the full set of terms in the VAD lexicon. Row *b.* shows the SHR on just the Warriner subset of terms in the VAD lexicon. Row *c.* shows the SHR reported by Warriner et al. (2013) on their annotations. Observe that the SHR scores for our annotations are markedly higher than those reported by Warriner et al. (2013), especially for arousal and dominance. All differences in SHR scores between rows b and c are statistically significant.

Summary of Main Results: The low correlations between the scores in our lexicon and the Warriner lexicon (especially for D and A) show that the scores in the two lexicons are substantially different. The scores for correlations across all pairs of dimensions in our lexicon are low ($r < 0.5$). SHR scores of 0.95 for valence, 0.9 for arousal, and 0.9 for dominance show for the first time that highly reliable fine-grained ratings can be obtained for valence, arousal, and dominance.

6 Shared Understanding of VAD Within and Across Demographic Groups

Human cognition and behaviour is impacted by evolutionary and socio-cultural factors. These factors are known to impact different groups of people differently (men vs. women, young vs. old, etc.). Thus it is not surprising that our understanding of the world may be slightly different depending on our demographic attributes. Consider gender—a key demographic attribute.[14] Men, women, and other genders are substantially more alike than they are different. However, they have encountered different socio-cultural influences for thousands of years. Often these disparities have been a means to exert unequal status and asymmetric power relations. Thus a crucial area in gender studies is to examine both the overt and subtle impacts of these socio-cultural influences, as well as ways to mitigate the inequity. Understanding how different genders perceive and use language is an important component of that research. Language use is also relevant to the understanding and treatment of neuropsychiatric disorders, such as sleep, mood, and anxiety disorders, which have been shown to occur more frequently in women than men (Bao and Swaab, 2011; Lewinsohn et al., 1998; McLean et al., 2011; Johnson et al., 2006; Chmielewski et al., 1995).

In addition to the VAD Lexicon (created by aggregating human judgments), we also make available the demographic information of the annotators. This demographic information along with the individual judgments on the best–worst tuples forms a significant resource in the study of how demographic attributes are correlated with our understanding of language. The data can be used to shed light on research questions such as: 'are there significant differences in the shared understanding of word meanings in men and women?', 'how is the social construct of gender reflected in language, especially in socio-political interactions?', 'does age impact our view of the valence, arousal, and dominance of concepts?', 'do people that view themselves as conscientious have slightly different judgments of valence, arousal, and dominance, than people who view themselves as easy going?', and so on.

[14] Note that the term *sex* refers to a biological attribute pertaining to the anatomy of one's reproductive system and sex chromosomes, whereas *gender* refers to a psycho-socio-cultural construct based on a person's sex or a person's self identification of levels of masculinity and femininity. One may identify their gender as female, male, agender, trans, queer, etc.

	V	A	D
f–f pairs	56.55	44.15	42.55
m–m pairs	56.88	43.80	43.55
f–m pairs	56.41	43.65	43.03

Table 7: Gender: Average agreement % on best–worst responses.

	V	A	D
f–f pairs vs. m–m pairs	y	y	y
f–f pairs vs. f–m pairs	-	y	y
m–m pairs vs. f–m pairs	y	-	y

Table 8: Gender: Significance of difference in average agreement scores (p = 0.05). 'y' = yes significant. '-' = not significant.

6.1 Experiments

We now describe experiments we conducted to determine whether demographic attributes impact how we judge words for valence, arousal, and dominance. For each demographic attribute, we partitioned the annotators into two groups: male (m) and female (f), ages 18 to 35 ($\leq$35) and ages over 35 (>35), and so on.[15] For each of the five personality traits, annotators are partitioned into the two groups shown in the bullet list of Section 4. We then calculated the extent to which people within the same group agreed with each other, and the extent to which people across groups agreed with each other on the VAD annotations (as described in the paragraph below). We also determined if the differences in agreement were statistically significant.

For each dimension (V, A, and D), we first collected only those 4-tuples where at least two female and at least two male responses were available. We will refer to this set as the *base set*. For each of the base set 4-tuples, we calculated three agreement percentages: 1. the percentage of all female–female best–worst responses where the two agreed with each other, 2. the percentage of all male–male responses where the two agreed with each other, and 3. the percentage of all female–male responses where the two agreed with each other. We then calculated the averages of the agreement percentages across all the 4-tuples in the base set. We conducted similar experiments for age groups and personality traits.

[15]For age, we chose 35 to create the two groups because several psychology and medical studies report changes in health and well-being at this age. Nonetheless, other partitions of age are also worth exploring.

	V	A	D
$\leq$35–$\leq$35 pairs	56.10	43.84	43.81
>35–>35 pairs	57.56	44.10	42.49
$\leq$35–>35 pairs	56.40	43.58	43.07

Table 9: Age: Average agreement % on best–worst responses.

	V	A	D
$\leq$35–$\leq$35 pairs vs. >35–>35 pairs	y	y	y
$\leq$35–$\leq$35 pairs vs. $\leq$35–>35 pairs	y	y	y
>35–>35 pairs vs. $\leq$35–>35 pairs	y	y	y

Table 10: Age: Significance of difference in average agreement scores (p = 0.05).

6.2 Results

Table 7 shows the results for gender. Note that the average agreement numbers are not expected to be high because often a 4-tuple may include two words that are close to each other in terms of the property of interest (V/A/D).[16] However, the relative values of the agreement percentages indicate the relative levels of agreements within groups and across groups.

Table 7 numbers indicate that women have a higher shared understanding of the degree of arousal of words (higher f–f average agreement scores on A), whereas men have a higher shared understanding of dominance and valence of words (higher m–m average agreement scores on V and D). The table also shows the cross-group (f–m) average agreements are the lowest for valence and arousal, but higher than f–f pairs for dominance. (Each of these agreements was determined from 1 to 1.5 million judgment pairs.)

Table 8 shows which of the Table 7 average agreements are statistically significantly different (shown with a 'y'). Significance values were calculated using the chi-square test for independence and significance level of 0.05. Observe that all score differences are statistically significant except for between f–f and f–m scores for V and m–m and f–m scores for A.

Tables 9 through 12 are similar to Tables 7 and 8, but for age groups and personality traits. Tables 9 and 10 show that respondents over the age of 35 obtain significantly higher agreements with each other on valence and arousal and lower agreements on dominance, than respondents aged 35 and under (with each other). Tables 11 and 12 show that

[16]Such disagreements are useful as they cause the two words to obtain scores close to each other.

	V	A	D
Agreeable (Ag) – Disagreeable (Di)			
# pairs	*1.0M*	*1.8M*	*1.7M*
Ag–Ag pairs	56.54	43.89	42.39
Di–Di pairs	55.76	43.63	43.61
Ag–Di pairs	56.28	43.57	43.01
Conscientious (Co) – Easygoing (Ea)			
# pairs	*0.9M*	*1.9M*	*1.5M*
Co–Co pairs	56.34	44.60	44.38
Ea–Ea pairs	56.39	43.15	41.36
Co–Ea pairs	56.39	43.77	42.52
Extrovert (Ex) – Introvert (In)			
# pairs	*0.9M*	*2.0M*	*1.6M*
Ex–Ex pairs	58.00	44.16	43.43
In–In pairs	56.49	43.78	42.16
Ex–In pairs	57.00	43.85	42.89
Neurotic (Ne) – Secure (Se)			
# pairs	*1.0M*	*1.8M*	*1.5M*
Ne–Ne pairs	56.33	43.78	41.98
Se–Se pairs	57.97	43.90	43.65
Ne–Se pairs	56.93	43.97	42.93
Open (Op) – Closed (Cl)			
# pairs	*0.8M*	*1.8M*	*1.3M*
Op–Op pairs	57.65	44.19	43.51
Cl–Cl pairs	56.39	43.52	43.23
Op–Cl pairs	56.90	44.03	43.36

Table 11: Personality Trait: Average agreement % on best–worst responses.

	V	A	D
Agreeable (Ag) – Disagreeable (Di)			
Ag–Ag vs. Di–Di	y	y	y
Ag–Ag vs. Ag–Di	y	y	y
Di–Di vs. Ag–Di	y	-	y
Conscientious (Co) – Easygoing (Ea)			
Co–Co vs. Ea–Ea	-	y	y
Co–Co vs. Co–Ea	-	y	y
Ea–Ea vs. Co–Ea	-	y	y
Extrovert (Ex) – Introvert (In)			
Ex–Ex vs. In–In	y	y	y
Ex–Ex vs. Ex–In	y	-	y
In–In vs. Ex–In	y	y	y
Neurotic (Ne) – Secure (Se)			
Ne–Ne vs. Se–Se	y	-	y
Ne–Ne vs. Ne–Se	y	-	y
Se–Se vs. Ne–Se	y	-	y
Open (Op) – Closed (Cl)			
Op–Op vs. Cl–Cl	y	y	y
Op–Op vs. Op–Cl	y	-	-
Cl–Cl vs. Op–Cl	y	y	-

Table 12: Personality Trait: Significance of difference in average agreement scores (p = 0.05).

some personality traits significantly impact a person's annotations of one or more of V, A, and D. Notably, those who view themselves as conscientious have a particularly higher shared understanding of the dominance of words, as compared to those who view themselves as easy going. They also have higher in-group agreement for arousal, than those who view themselves as easy going, but the difference for valence is not statistically significant. Also notable, is that those who view themselves as extroverts have a particularly higher shared understanding of the valence, arousal, and dominance of words, as compared to those who view themselves as introverts.

Finally, as a sanity check, we divided respondents into those whose CrowdFlower worker ids are odd and those whose worker ids are even. We then determined average agreements for even–even, odd-odd, and even–odd groups just as we did for the demographic variables. We found that, as expected, there were no significant differences in average agreements.

Summary of Main Results: We showed that several demographic attributes such as age, gender, and personality traits impact how we judge words for valence, arousal, and dominance. Further, people that share certain demographic attributes show a higher shared understanding of the relative rankings of words by (one or more of) V, A, or D than others. However, this raises new questions: why do certain demographic attributes impact our judgments of V, A, and D? Are there evolutionary forces that caused some groups such as women to develop a higher shared understanding or the arousal, whereas different evolutionary forces caused some groups, such as men, to have a higher shared understanding of dominance? We hope that the data collected as part of this project will spur further inquiry into these and other questions.

7 Applications and Future Work

The large number of entries in the VAD Lexicon and the high reliability of the scores make it useful for a number of research projects and applications. We list a few below:

- To provide features for sentiment or emotion detection systems. They can also be used to obtain sentiment-aware word embeddings and sentiment-aware sentence representations.
- To study the interplay between the basic emotion model and the VAD model of affect. The VAD lexicon can be used along with lists of words associated with emotions such as joy, sadness, fear, etc. to study the correlation of V, A, and D, with those emotions.

- To study the role emotion words play in high emotion intensity sentences or tweets. The Tweet Emotion Intensity Dataset has emotion intensity and valence scores for whole tweets (Mohammad and Bravo-Marquez, 2017). We will use the VAD lexicon to determine the extent to which high intensity and high valence tweets consist of high V, A, and D words, and to identify sentences that express high emotional intensity without using high V, A, and D words.
- To identify syllables that tend to occur in words with high VAD scores, which in turn can be used to generate names for literary characters and commercial products that have the desired affectual response.
- To identify high V, A, and D words in books and literature. To facilitate research in digital humanities. To facilitate work on literary analysis.
- As a source of gold (reference) scores, the entries in the VAD lexicon can be used in the evaluation of automatic methods of determining V, A, and D.
- To analyze V, A, ad D annotations for different groups of words, such as: hashtag words and emojis common in tweets, emotion denotating words, emotion associated words, neutral terms, words belonging to particular parts of speech such as nouns, verbs, and adjectives, etc.
- To analyze interactions between demographic groups and specific groups of words, for example, whether younger annotators have a higher shared understanding of tweet terms, whether a certain gender is associated with a higher shared understanding of adjectives, etc.
- To analyze the shared understanding of V, A, and D within and across geographic and language groups. We are interested in creating VAD lexicons for other languages. We can then explore characteristics of valence, arousal, and dominance that are common across cultures. We can also test whether some of the conclusions reached in this work apply only to English, or more broadly to multiple languages.
- The dataset is of use to psychologists and evolutionary linguists interested in determining how evolution shaped our representation of the world around us, and why certain personality traits are associated with higher or lower shared understanding of V, A, and D.

8 Conclusions

We obtained reliable human ratings of valence, arousal, and dominance for more than 20,000 English words. (It has about 40% more words than the largest existing manually created VAD lexicon). We used best–worst scaling to obtain fine-grained scores (and word rankings) and addressed issues of annotation consistency that plague traditional rating scale methods of annotation. We showed that the lexicon has split-half reliability scores of 0.95 for valence, 0.90 for arousal, and 0.90 for dominance. These scores are markedly higher than that of existing lexicons.

We analyzed demographic information to show that even though the annotations overall lead to consistent scores in repeated annotations, there exist statistically significant differences in agreements across demographic groups such as males and females, those above the age of 35 and those that are 35 or under, and across personality dimensions (extroverts and introverts, neurotic and secure, etc.). These results show that certain demographic attributes impact how we view the world around us in terms of the relative valence, arousal, and dominance of the concepts in it.

The NRC Valence, Arousal, and Dominance Lexicon is made available.[17] It can be used in combination with other manually created affect lexicons such as the NRC Word–Emotion Association Lexicon (Mohammad and Turney, 2013)[18] and the NRC Affect Intensity Lexicon (Mohammad, 2018).[19]

Acknowledgments

Many thanks to Svetlana Kiritchenko, Michael Wojatzki, Norm Vinson, and Tara Small for helpful discussions.

[17] **The NRC Valence, Arousal, and Dominance Lexicon** provides human ratings of valence, arousal, and dominance for more than 20,000 English words:
http://saifmohammad.com/WebPages/nrc-vad.html

[18] **The NRC Emotion Lexicon** includes about 14,000 words annotated to indicate whether they are associated with any of the eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust):
http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[19] **The NRC Affect Intensity Lexicon** provides real-valued affect intensity scores for four basic emotions (anger, fear, sadness, joy):
http://saifmohammad.com/WebPages/AffectIntensity.htm

References

Iris Bakker, Theo van der Voordt, Peter Vink, and Jan de Boon. 2014. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3):405–421.

Ai-Min Bao and Dick F Swaab. 2011. Sexual differentiation of the human brain: relation to gender identity, sexual orientation and neuropsychiatric disorders. *Frontiers in neuroendocrinology*, 32(2):214–226.

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26.

Hans Baumgartner and Jan-Benedict E.M. Steenkamp. 2001. Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2):143–156.

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.

Phillip M Chmielewski, Leyan OL Fernandes, Cindy M Yee, and Gregory A Miller. 1995. Ethnicity and gender in scales of psychosis proneness and mood disorders. *Journal of Abnormal Psychology*, 104(3):464.

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.

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*, pages 178–201. Edward Elgar Publishing.

Eric O Johnson, Thomas Roth, Lonni Schultz, and Naomi Breslau. 2006. Epidemiology of dsm-iv insomnia in adolescence: lifetime prevalence, chronicity, and an emergent gender difference. *Pediatrics*, 117(2):e247–e256.

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*, pages 356–364, Montréal, Canada.

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.

Svetlana Kiritchenko and Saif M. Mohammad. 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.

Peer M Lewinsohn, Ian H Gotlib, Mark Lewinsohn, John R Seeley, and Nicholas B Allen. 1998. Gender differences in anxiety disorders and anxiety symptoms in adolescents. *Journal of abnormal psychology*, 107(1):109.

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.

Carmen P McLean, Anu Asnaani, Brett T Litz, and Stefan G Hofmann. 2011. Gender differences in anxiety disorders: prevalence, course of illness, comorbidity and burden of illness. *Journal of psychiatric research*, 45(8):1027–1035.

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

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. 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.

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

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

Saif M. Mohammad and Svetlana Kiritchenko. 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.

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

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 dutch words. *Behavior research methods*, 45(1):169–177.

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

C.E. Osgood, Suci G., and P. Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

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

Stanley Presser and Howard Schuman. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. SAGE Publications, Inc.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

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

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

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.

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

Melissa LH Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J Hofmann, and Arthur M Jacobs. 2009. The berlin affective word list reloaded (bawl-r). *Behavior research methods*, 41(2):534–538.

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.

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.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xue-jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 788–793.