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 @SaifMMohammad

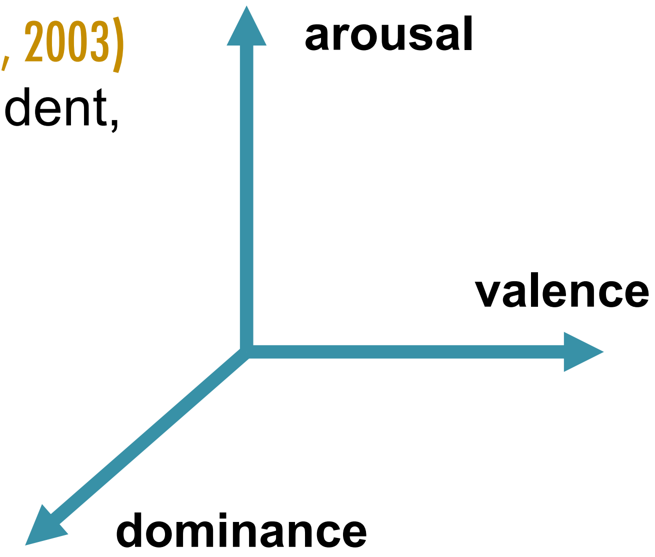
National Research Council Canada Conseil national de recherches Canada

Canada

Core Dimensions of Meaning

Influential factor analysis studies (Osgood et al., 1957; Russell, 1980, 2003) have shown that the three most important, largely independent, dimensions of word meaning:

- valence (V): positive/pleasure – negative/displeasure
- arousal (A): active/stimulated – sluggish/bored
- dominance (D): powerful/strong – powerless/weak



Thus, when comparing the meanings of two words, we can compare their V, A, D scores. For example:

- *banquet* indicates more positiveness than *funeral*
- *nervous* indicates more arousal than *lazy*
- *queen* indicates more dominance than *delicate*

This work:

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

fine-grained

This work:

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance

for 20,000 English Words

Motivation

Human annotations of words for VAD

- For use by automatic systems:
  - predicting VAD of words
  - predicting sentiment and emotions of sentences, tweets, etc.
  - detecting stance, personality traits, well-being, cyber-bullying, etc.

- To draw inferences about people:
  - to understand how we (or different groups of people) use language to express meaning and emotions
    - analyze text written/spoken by different groups of people
    - analyze VAD judgments of different groups of people

Related Work: Existing VAD Lexicons

Affective Norms of English Words (ANEW) (Bradley and Lang, 1999)

- ~1,000 words
- 9-point rating scale

Warriner et al. Norms (Warriner et al. 2013)

- 14,000 words
- 9-point rating scale

Small number of VAD lexicons in non-English languages as well

- E.g.:
  - Moors et al. (2013) for Dutch
  - Vo et al. (2009) for German
  - Redondo et al. (2007) for Spanish
- rating scale

Related Work: Existing VAD Lexicons

Affective Norms of English Words (ANEW) (Bradley and Lang, 1999)

- ~1,000 words
- 9-point rating scale

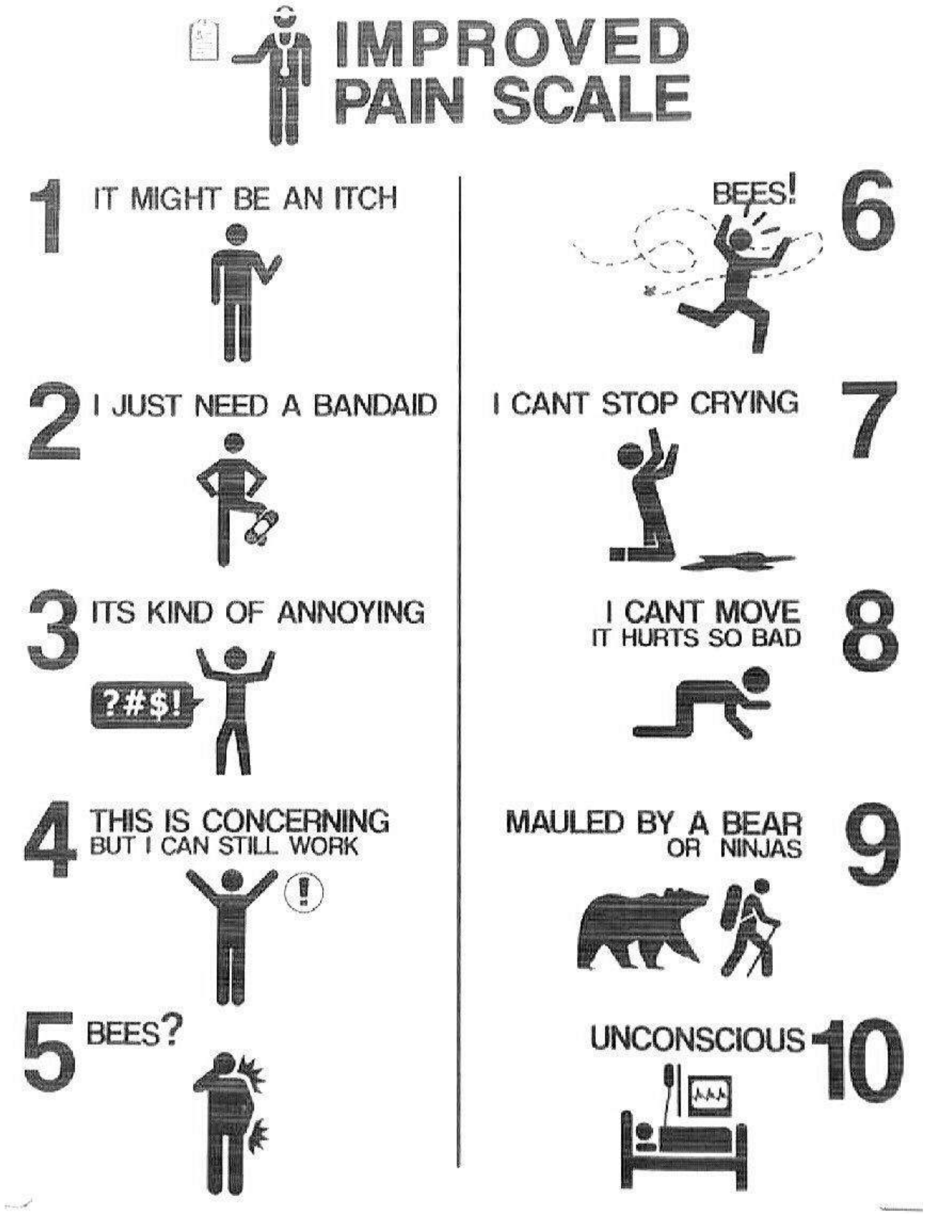
Warriner et al. Norms (Warriner et al. 2013)

- 14,000 words
- 9-point rating scale

Small number of VAD lexicons in non-English languages as well

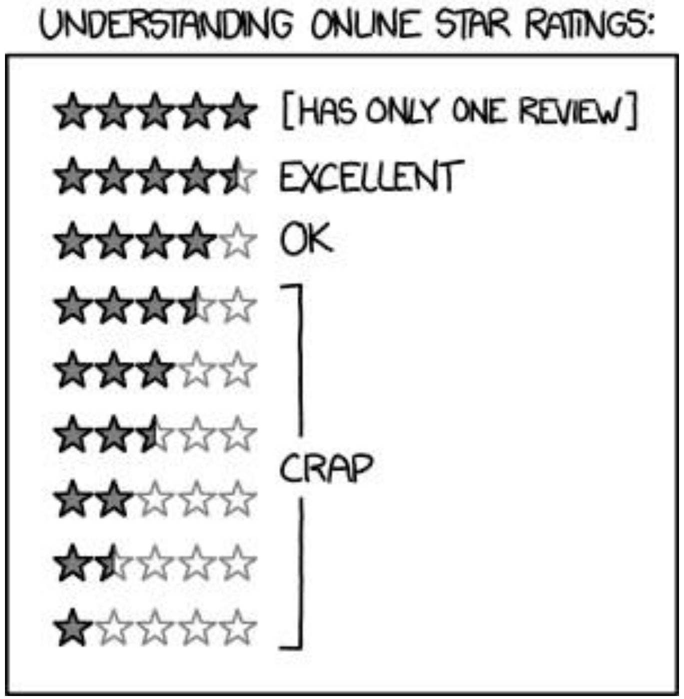
- E.g.:
    - Moors et al. (2013) for Dutch
    - Vo et al. (2009) for German
    - Redondo et al. (2007) for Spanish
- rating scale

Rating scales:



source: imgur

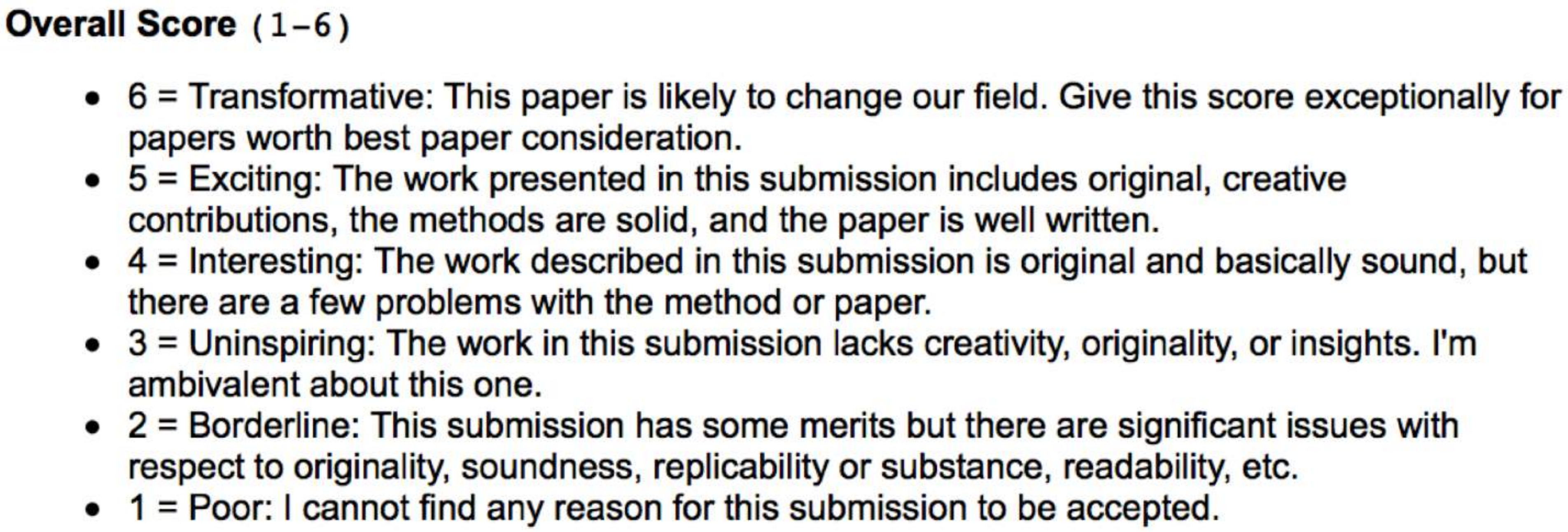
Rating scales:


UNDERSTANDING ONLINE STAR RATINGS:
★★★★★ [HAS ONLY ONE REVIEW]
EXCELLENT
OK
CRAP

source: xkcd

Rating scales:

ACL-2018 Reviewing Scale

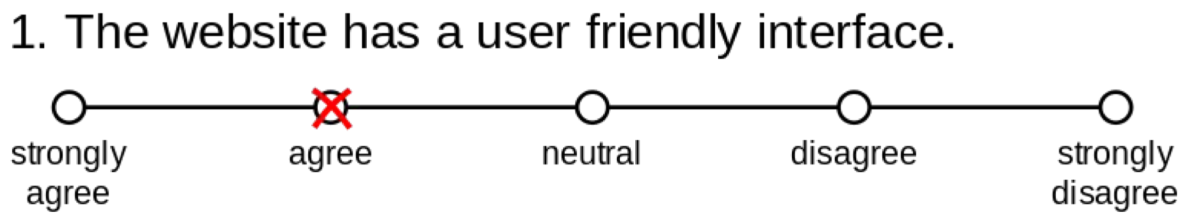


Overall Score (1-6)

- 6 = Transformative: This paper is likely to change our field. Give this score exceptionally for papers worth best paper consideration.
- 5 = Exciting: The work presented in this submission includes original, creative contributions, the methods are solid, and the paper is well written.
- 4 = Interesting: The work described in this submission is original and basically sound, but there are a few problems with the method or paper.
- 3 = Uninspiring: The work in this submission lacks creativity, originality, or insights. I'm ambivalent about this one.
- 2 = Borderline: This submission has some merits but there are significant issues with respect to originality, soundness, replicability or substance, readability, etc.
- 1 = Poor: I cannot find any reason for this submission to be accepted.

Rating scales:

Likert Item (Likert 1932)

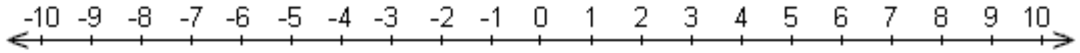


1. The website has a user friendly interface.

○ strongly agree — [✓ agree] — ○ neutral — ○ disagree — ○ strongly disagree

source: Wikimedia Commons

Note: A Likert scale is the sum of responses on several Likert items.



Problems with rating scales:

- fixed granularity
- difficult to maintain consistency across annotators
- difficult for an annotator to be self consistent
- scale region bias

Comparative Annotations

Paired Comparisons (Thurstone, 1927; David, 1963)**:**

If X is the property of interest (positive, useful, etc.), give two terms and ask which is more X

- less cognitive load
- helps with consistency issues
- requires a large number of annotations
  - order N^2, where N is number of terms to be annotated

Best–Worst Scaling (BWS) (Louviere & Woodworth, 1990)

- The annotator is presented with four words (say, A, B, C, and D) and asked:
  - which word is associated with the most/highest X (property of interest, say valence)
  - which word is associated with the least/lowest X

- By answering just these two questions, five out of the six inequalities are known
  - For e.g.:
    - If A: highest valence
    - and D: lowest valence, then we know:

A > B, A > C, A > D, B > D, C > D

Best–Worst Scaling (Louviere & Woodworth, 1990)

- Each of these BWS questions can be presented to multiple annotators.
- We can obtain real-valued scores for all the terms using a simple counting method (Orme, 2009)

  score(w) = (#best(w) - #worst(w)) / #annotations(w)

  the scores range from:
  -1 (least X)
  to 1 (most X)

  X = property of interest, say valence

  - the scores can then be used to rank all the terms

Best–Worst Scaling (Louviere & Woodworth, 1990)

- preserves the comparative nature
- keeps the number of annotations down to about 2N
- leads to more reliable, less biased, more discriminating annotations (Kiritchenko and Mohammad, 2017, Cohen, 2003)

Creating the Valence, Arousal, and Dominance Lexicon

Term Selection

We wanted to include:

- commonly used English terms
- terms common in tweets
- terms that denotate or connotate emotions

Selected:

- All terms in the NRC Emotion Lexicon (Mohammad and Turney, 2013): ~14,000
  - labels indicate association with eight basic emotions
    anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Plutchik, 1980)
  - includes terms that occur frequently in the Google n-gram corpus
- All terms in ANEW (Bradley and Lang, 1999): ~1000
- All terms in the Warriner et al. lexicon (2013): ~14,000
- Words from the Roget's Thesaurus categories corresponding to the eight basic Plutchik emotions: ~520
- High-frequency content terms, including emoticons, from the Hashtag Emotion Corpus (a tweets corpus) (Mohammad, 2012): ~1000

Total: 20,007 terms

Best-Worst Questionnaires

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)

Similar questions for arousal and dominance

This study was approved by the NRC Research Ethics Board (NRC-REB) under protocol number 2017-98.
REB review seeks to ensure that research projects involving humans as participants meet Canadian standards of ethics.

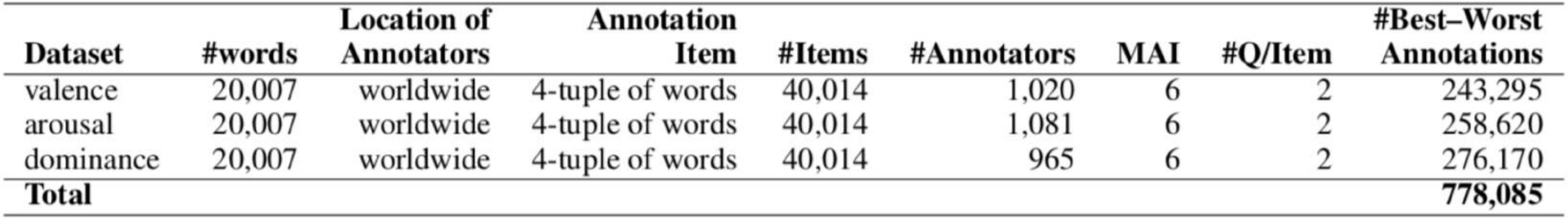
Crowdsourcing and Quality Control

About 2% of the data was annotated internally beforehand (by the author)

- These gold questions are interspersed with other questions
- If one gets a gold question wrong, they are immediately notified of it
  - feedback to improve task understanding
- If one's accuracy on the gold questions falls below 80%,
  - they are refused further annotation
  - all of their annotations are discarded

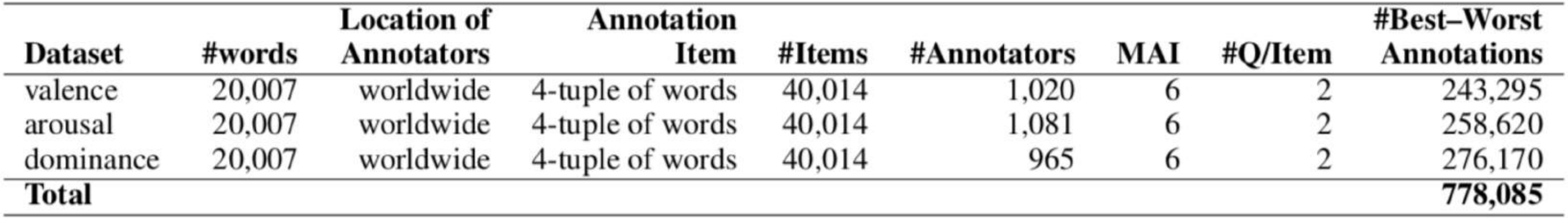
Mechanism to avoid malicious or random annotations

Valence, Arousal, and Dominance Annotations (with BWS)



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**

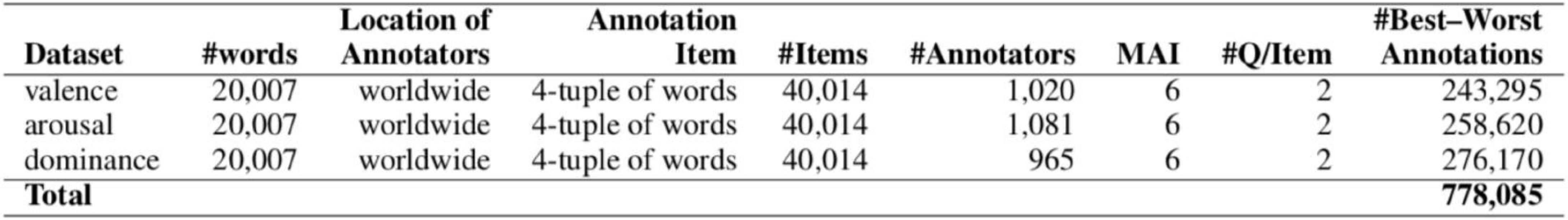
Valence, Arousal, and Dominance Annotations (with BWS)



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**

2N 4-tuples

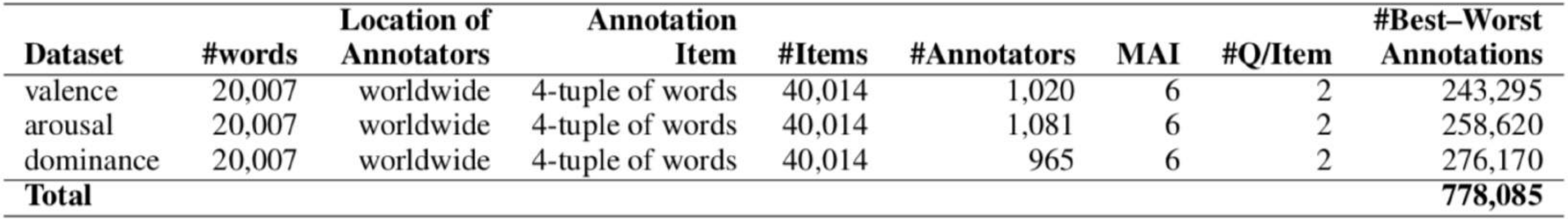
Valence, Arousal, and Dominance Annotations (with BWS)



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**

~1000 annotators for each task

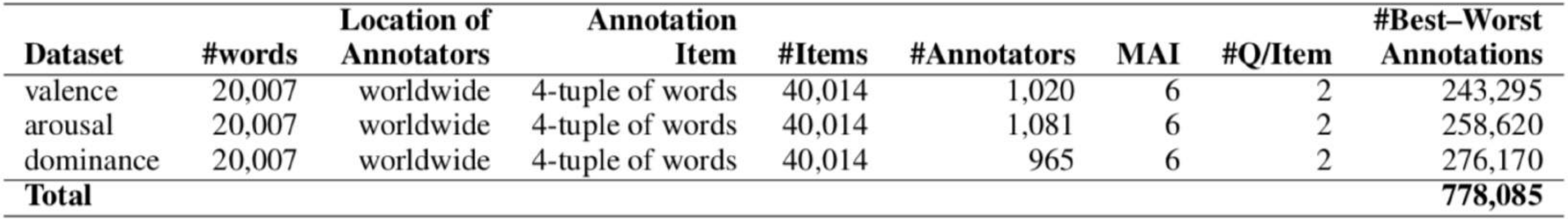
Valence, Arousal, and Dominance Annotations (with BWS)



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**

minimum and median annotations per 4-tuple

Valence, Arousal, and Dominance Annotations (with BWS)



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**

number of pairs of best—worst annotations

Best–Worst Scaling (Louviere & Woodworth, 1990)

- Each of these BWS questions can be presented to multiple annotators.
- We can obtain real-valued scores for all the terms using a simple counting method (Orme, 2009)

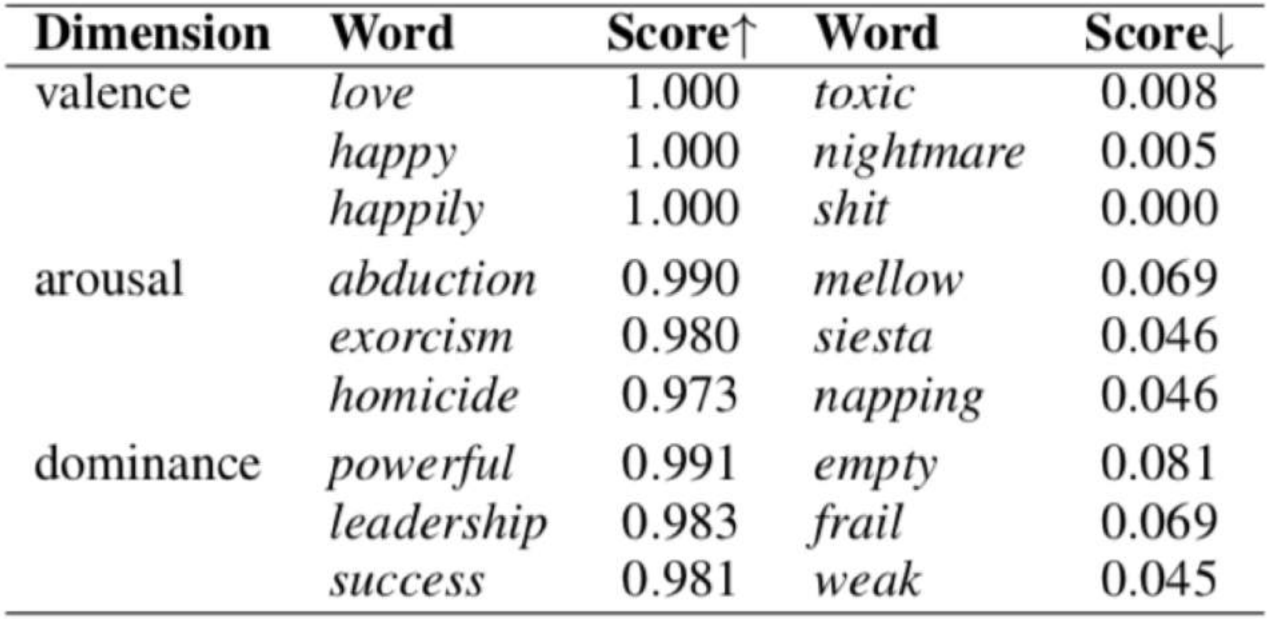
  score(w) = (#best(w) - #worst(w)) / #annotations(w)

  the scores range from:
  -1 (least X)
  to 1 (most X)

  X = property of interest, say valence

  - the scores can then be used to rank all the terms

Example Entries in the VAD Lexicon

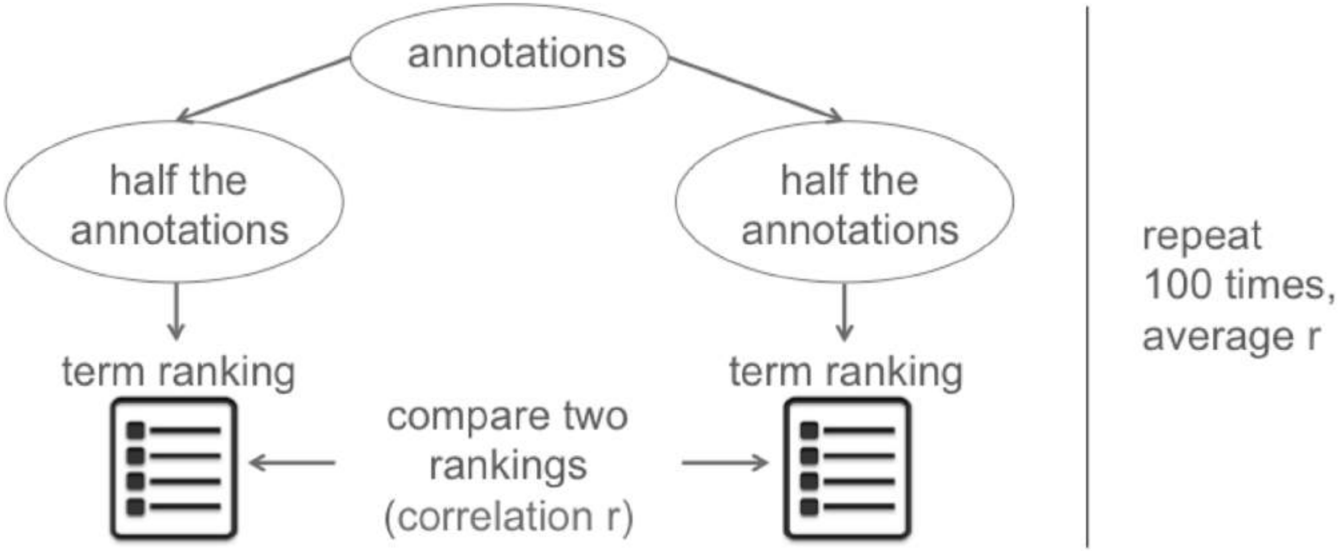


Dimension	Word	Score↑	Word	Score↓
valence	*love*	1.000	*toxic*	0.008
	happy	1.000	*nightmare*	0.005
	happily	1.000	*shit*	0.000
arousal	*abduction*	0.990	*mellow*	0.069
	exorcism	0.980	*siesta*	0.046
	homicide	0.973	*napping*	0.046
dominance	*powerful*	0.991	*empty*	0.081
	leadership	0.983	*frail*	0.069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

Scores are in the range 0 (lowest V/A/D) to 1 (highest V/A/D)

Reliability (Reproducibility) of Annotations

Average split-half reliability (SHR): a commonly used approach to determine consistency (Kuder and Richardson, 1937; Cronbach, 1946)



Pearson correlation: -1(most inversely correlated) to 1(most correlated)

Split-Half Reliability Scores for VAD Annotations

Annotations	# Terms	# Annotations	V	A	D
Warriner et al. (2013)	13,915	20 per term	0.91	0.79	0.77

Markedly lower SHR for A and D.
The dominance ratings seem especially problematic since the Warriner V-D correlation is 0.71.

Split-Half Reliability Scores for VAD Annotations

Annotations	# Terms	# Annotations	V	A	D
Warriner et al. (2013)	13,915	20 per term	0.91	0.79	0.77
Ours (Warriner terms)	13,915	6 per tuple	0.95	0.91	0.91

Split-Half Reliability Scores for VAD Annotations

Annotations	# Terms	# Annotations	V	A	D
Warriner et al. (2013)	13,915	20 per term	0.91	0.79	0.77
Ours (Warriner terms)	13,915	6 per tuple	0.95	0.91	0.91
Ours (all terms)	20,007	6 per tuple	0.95	0.90	0.90

These SHR scores show for the first time that highly reliable fine-grained ratings can be obtained for valence, arousal, and dominance. Also, our V-D correlation is 0.48.

NRC VAD Lexicon and the Warriner et al. Lexicon:

How Different are the Scores?

Pearson correlations r

Annotations	V	A	D
Ours-Warriner (for overlapping terms)	0.81	0.62	0.33

The especially low correlations for dominance and arousal indicate that our lexicon has substantially different scores and rankings of terms.

Done:

Create the large and reliable VAD lexicon

On to:

Analyze VAD judgments of different groups of people

Shared Understanding of VAD:

Within and Across Demographic Groups

- Human cognition and behaviour are impacted by evolutionary and socio-cultural factors
- These factors impact different groups of people differently
- Consider gender
    - Men, women, and other genders are substantially more alike than different
    - However, they have encountered different socio-cultural influences
    - Often these disparities have been a means to exert unequal status and asymmetric power relations
    - Gender studies examine
        - both the overt and subtle impacts of these socio-cultural influences
        - ways to mitigate the inequity
        - how different genders perceive and use language

Demographic Survey

Annotators could optionally respond to a separate survey asking for their demographic information:

- age
- gender
- country
- personality traits
  - we asked how they viewed themselves across the big five personality traits (Barrick and Mount, 1991)

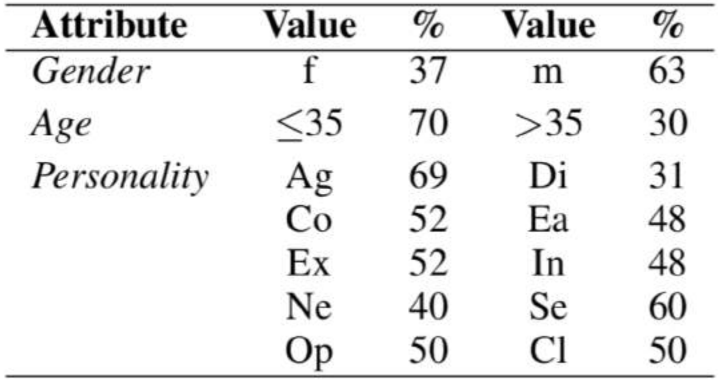
991 people (55% of the VAD annotators) chose to provide their demographic information

Experiment

- For each demographic attribute, we partitioned the annotators into two groups:
  - male (m) and female (f)
  - those 18 to 35 (≤35) and those over 35 (>35)

Experiment

- For each demographic attribute, we partitioned the annotators into two groups:
  - male (m) and female (f)
  - those 18 to 35 (young) and those over 35 (grownups)
  - agreeable (Ag) and Disagreeable (Di)
  - extrovert (Ex) and introvert (In)
  - and so on



Attribute	**Value**	**%**	**Value**	**%**
Gender	f	37	m	63
Age	$\leq$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

- Calculated
  - the extent to which people within the same group agreed with each other on the VAD annotations
  - whether the differences in average agreements in each group are significant
    - chi-square test for independence and significance level of 0.05

Differences in Average Agreements: Gender

Sub-group with Significantly Higher Agreement

	Valence	Arousal	Dominance
F–F vs. M–M			

F = female
M = male

Differences in Average Agreements: Gender

Sub-group with Significantly Higher Agreement

	Valence	Arousal	Dominance
F–F vs. M–M	M–M	F–F	M–M

F = female
M = male

Women have a higher shared understanding of the degree of arousal of words.
Men have a higher shared understanding of the dominance and valence of words.

Differences in Average Agreements: Age

	Sub-group with Significantly Higher Agreement		
	Valence	Arousal	Dominance
Y–Y vs. G–G			

Y = young
G = grownups

Differences in Average Agreements: Age

Sub-group with Significantly Higher Agreement

	Valence	Arousal	Dominance
Y–Y vs. G–G	G–G	G–G	Y–Y

Y = young
G = grownups

The young have a higher shared understanding of the dominance of words.
The grownups have a higher shared understanding of valence and arousal of words.

Differences in Average Agreements: Big 5 Traits

Sub-group with Significantly Higher Agreement

	Valence	Arousal	Dominance
Ag—Ag vs. Di—Di	Ag—Ag	Ag—Ag	Di—Di
Co—Co vs. Ea—Ea	-	Co—Co	Co—Co
Ex—Ex vs. In—In	Ex—Ex	Ex—Ex	Ex—Ex
Ne—Ne vs Se—Se	Se—Se	-	Se—Se
Op—Op vs Cl—Cl	Op—Op	Op—Op	Op—Op

Ag = Agreeableness (friendly and compassionate)
Di = Disagreeableness (careful in whom to trust, argumentative)

Co = Conscientiousness (efficient and organized)
Ea = Easygoing (easy-going and carefree)

Ex = Extrovert (outgoing, energetic, seek the company of others)
In = Introvert (solitary, reserved, meeting many people causes anxiety)

Ne = Neurotic (often feel anger, anxiety, depression, and vulnerability)
Se = Secure (rarely feel anger, anxiety, depression, and vulnerability)

Op = Open to experiences (inventive and curious; seek out new experiences)
Cl = Closed to experiences (consistent and cautious; anxious about new experiences)

Done:

Create the large and reliable VAD lexicon
Analyze VAD judgments of different groups of people

On to:

Applications and Summary

Selected Applications and Future Work

- Source of features for systems in sentiment, emotion, and other affect-related tasks
  - useful to create emotion-aware word embeddings and emotion-aware sentence representations
- Source of gold (reference) scores, to evaluate automatic methods of determining V, A, and D
- Study the interplay between the basic emotion model and the VAD model of emotions (Mohammad, 2018: LREC paper)
  - Companion lexicon: NRC Emotion Intensity Lexicon
    provides real-valued affect intensity scores for ~6000 words with four basic emotions (anger, fear, sadness, joy)
- Study the role of high VAD words in high emotion intensity sentences, tweets, snippets from literature

Summary

- Created the NRC Valence, Arousal, and Dominance Lexicon:
  - has entries for about 20,000 English words
  - has fine-grained real-valued scores for V, A, and D (core dimensions of meaning)
  - showed that the annotations are reliable (high split-half reliability scores)
- Showed that certain demographic attributes impact how we view the world around us.

The VAD lexicon is useful in a wide range of applications and research projects.

The NRC Valence, Arousal, and Dominance Lexicon

provides ratings of valence, arousal, and dominance for ~20,000 English words

http://saifmohammad.com/WebPages/nrc-vad.html

The NRC Word–Emotion Association Lexicon aka NRC Emotion Lexicon

provides associations for ~14,000 words with eight emotions (anger, fear, joy, sadness, anticipation, disgust, surprise, trust)

http://saifmohammad.com/WebPages/NRC-Emotion- Lexicon.htm

The NRC Emotion Intensity Lexicon aka Affect Intensity Lexicon

provides intensity scores for ~6000 words with four emotions (anger, fear, joy, sadness)

http://saifmohammad.com/WebPages/AffectIntensity.htm

The NRC Word–Colour Association Lexicon

provides associations for ~14,000 words with 11 common colours

http://saifmohammad.com/WebPages/lexicons.html

Pictures Attribution

Family by b farias from the Noun Project
Shovel and Pitchfork by Symbolon from the Noun Project
Checklist by Nick Bluth from the Noun Project
Generation by Creative Mahira from the Noun Project
Human by Adrien Coquet from the Noun Project
Search by Maxim Kulikov from the Noun Project

https://thenounproject.com

Resources Available at: www.saifmohammad.com

- NRC Valence, Arousal, and Dominance Lexicon
- NRC Emotion Lexicon and Emotion Intensity Lexicon
- Interactive visualizations

Saif M. Mohammad

✉ Saif.Mohammad@nrc-cnrc.gc.ca

@SaifMMohammad

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