



# Introduction:

## Word–Sentiment Associations

**Sentiment lexicon:** a list of terms (usually single words) with association to positive (negative) sentiment

happy 0.9

awful -0.9

award 0.6

Applications:

- sentence-, tweet-, message-level sentiment classification
- literary analysis
- detecting personality traits

**Our goal:** Manually capture fine-grained (real-valued) sentiment associations for single words and multi-word phrases

# Motivation:

## Manually Obtained Sentiment Annotations

- Manually created lexicons are generally more accurate than automatically generated lexicons
- Uses (that cannot be fulfilled by automatic lexicons):
  - to create automatic lexicons
  - to directly evaluate automatic lexicons
  - linguistic analysis
    - help understand how sentiment is conveyed by words and phrases
    - how sentiment is perceived by native speakers

# Motivation:

## Fine-Grained Sentiment Annotations

Existing manually created lexicons:

- usually have only coarse levels of sentiment (positive vs. negative)

Obtaining real-valued sentiment annotations is challenging:

- higher cognitive load than simply marking positive, negative, neutral
- hard to be consistent across multiple annotations
- difficult to maintain consistency across annotators
  - 0.8 for one annotator may be 0.7 for another

# Our Contributions

- Investigate the applicability and reliability of **Best–Worst Scaling** in sentiment annotation via crowdsourcing
- Create new **fine-grained sentiment lexicons** through manual annotation and Best–Worst Scaling
  - for different domains and languages
  - for words and also for phrases
- Show that the annotation method we use produces **reliable sentiment scores** with just two or three annotations per question
- Analyze the lexicons to gain new understandings of human perception of sentiment

# Annotation Method

**Best–Worst Scaling** (Louviere & Woodworth, 1990):  
(a.k.a. Maximum Difference Scaling or MaxDiff)

If X is the property of interest (positive, useful, etc.),

give k terms (usually 4 or 5) and ask which is most X, and which is least X



- **comparative** in nature
- **helps with consistency** issues

## Crowdsourcing:

- Each 4-tuple is annotated by at least eight respondents

# Best–Worst Scaling:

## Converting Responses to Real-Valued Scores

- Responses converted into real-valued scores for all the terms:
  - a simple counting procedure (Orme, 2009):

$$score(t) = \frac{\#most\ positive(t) - \#most\ negative(t)}{\#annotations(t)}$$

The scores range from:

-1 (least association with positive sentiment)  
to 1 (most association with positive sentiment)

- terms can then be ranked by sentiment

# New, Manually Created, Sentiment Lexicons

- We created three fine-grained sentiment lexicons:
  - **SemEval-2015 English Twitter**
    - 1,515 single words and negated phrases from English tweets (e.g., *happieee*, *can't wait*, *lmao*, *<33*)
  - **SemEval-2016 Arabic Twitter**
    - 1,367 single words and negated phrases from Arabic tweets (e.g., *صداءع*, *مش هيتحقق*, *#عشق*, *كارث*)
  - **SemEval-2016 General English Sentiment Modifiers** (aka **Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs**)
    - 3,207 single words and phrases with negators, modals, and degree adverbs (e.g., *delightful*, *rather dangerous*, *may not know*)

# Robustness of the Annotations

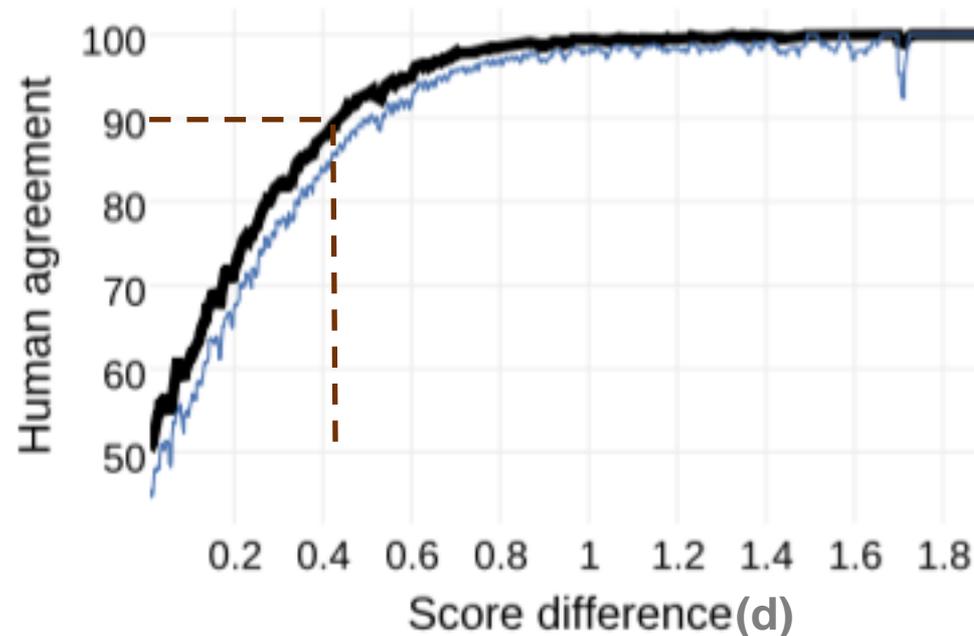
- Divided the Best–Worst responses for each question into two halves
- Generated scores and rankings based on each set individually
- The two sets produced very similar results:
  - Spearman Rank Correlation coefficient between the two rankings was 0.98 for all three lexicons
  - Pearson Correlation coefficient between the two sets of scores was 0.98 for all three lexicons



# Analysis:

## Human Agreement vs. Sentiment Difference

- For word pair  $w_1$  and  $w_2$  such that  $\text{score}(w_1) > \text{score}(w_2)$ , we calculate human agreement for  $\text{score}(w_1) > \text{score}(w_2)$
- We plot average human agreement as a function of  $d = \text{score}(w_1) - \text{score}(w_2)$



# Analysis:

## Least Perceptible Difference

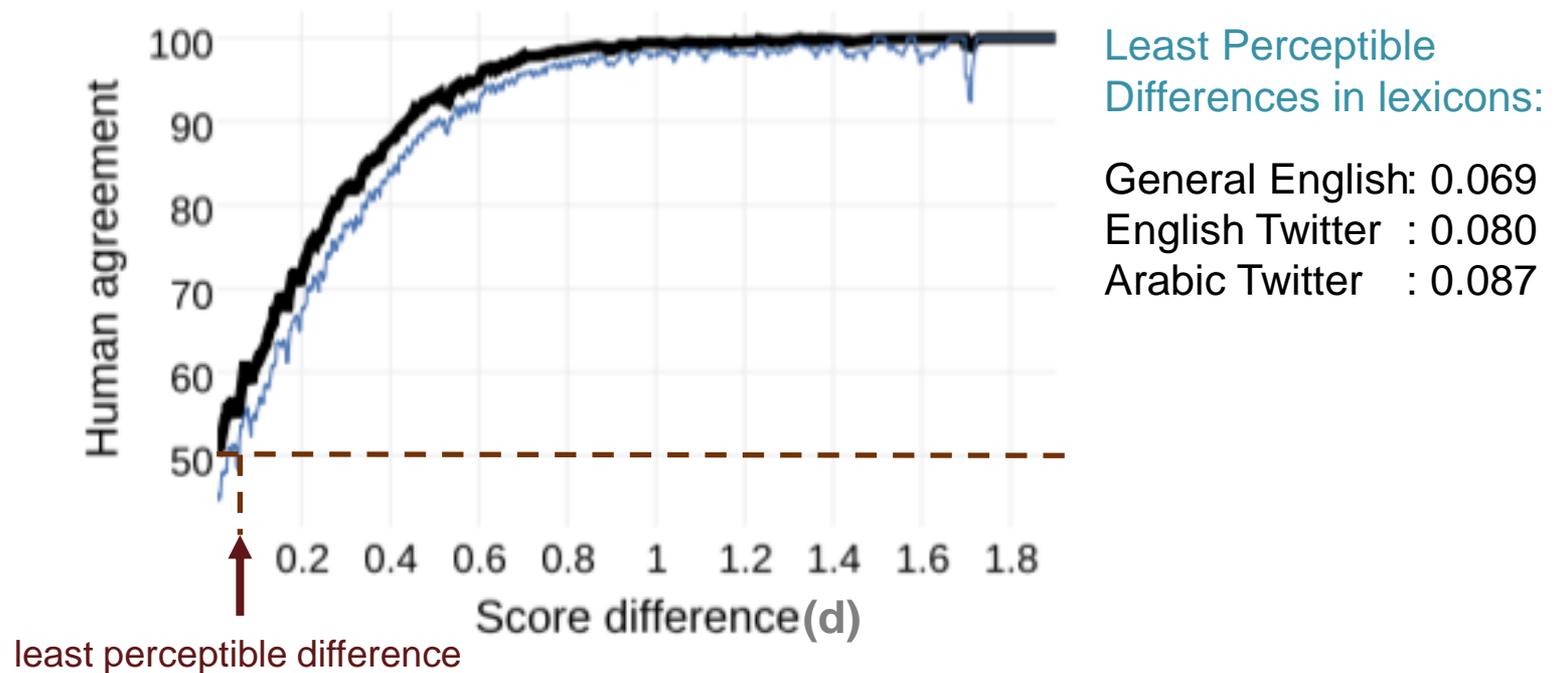


- Least perceptible difference aka just-noticeable difference
  - a concept from psychophysics
  - the amount by which something that can be measured (e.g., weight or sound intensity) needs to be changed in order for the difference to be noticeable by a human (Fechner, 1966)
- With our fine-grained sentiment scores, we can measure the least perceptible difference in sentiment
  - useful in studying sentiment composition (e.g., to determine whether a modifier significantly impacts the sentiment of the word it modifies)

# Analysis:

## Measuring the Least Perceptible Difference

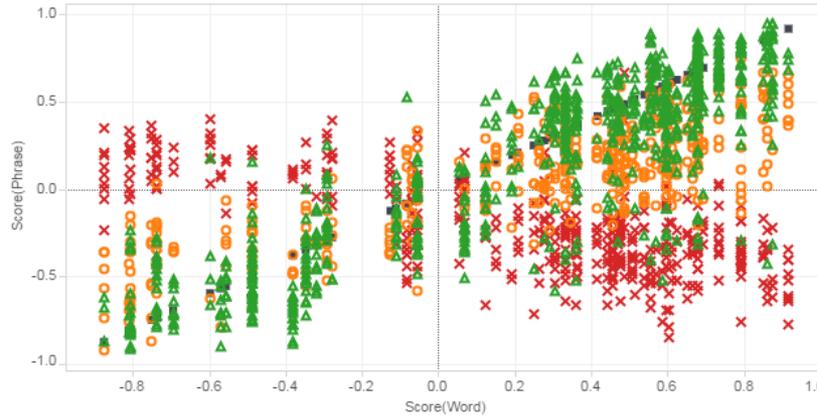
- Least perceptible difference in sentiment scores is a point  $d$  at which we can say with high confidence that the two terms do not have the same sentiment associations



# Interactive Visualization for SCL-NMA

Sentiment of a word vs. Sentiment of phrases consisting that word

Compressed x axis (sentiment of word)



Modifier Class

- ▲ adverb
- modal
- × negator
- word

Modifier Class

- adverb
- modal
- negator
- word

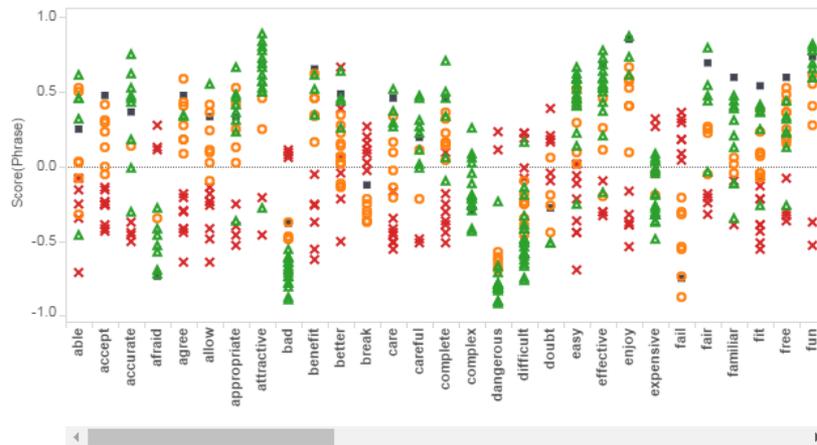
Modifier Class

- (All)
- adverb
- modal
- negator
- word

Modifier Word/Phrase

- (All)
- Null
- can
- can be
- cannot
- certainly
- could
- could be
- could not
- did not
- does not
- especially
- extremely
- fairly
- had no
- have no
- highly
- increasingly
- less
- may
- may be

Expanded x axis (sentiment of word)



Score(Phrase)

-0.921 0.944

<http://www.saifmohammad.com/WebPages/SCL.html#NMA>



# Lexicons Availability



The lexicons and their interactive visualizations are available at:  
<http://www.saifmohammad.com/WebPages/SCL.html>

Code for Best–Worst Scaling will be available at:  
<http://www.saifmohammad.com/WebPages/BestWorst.html>

The datasets were used as official test sets in:

- **SemEval-2015 Task 10**: English Twitter dataset  
<http://alt.qcri.org/semEval2015/task10/>
- **SemEval-2016 Task 7**: General English and Arabic Twitter datasets  
<http://alt.qcri.org/semEval2016/task7/>

We hope you will use **Best–Worst Scaling** for your next annotation project!