The Search for Emotions, Creativity, and Fairness in Language

Saif M. Mohammad
Senior Research Scientist, National Research Council Canada

Saif.Mohammad@nrc-cnrc.gc.ca  @SaifMMohammad
Emotions

- Determine human experience and behavior
- Condition our actions
- Central in organizing meaning
  - No cognition without emotion
The Search for Emotions in Language

creativity

fairness
How many emotions can we perceive?

Difficult question:
- fuzzy emotion boundaries, overlapping meanings, socio-cultural influences, etc.

Some studies suggest 500 to 600 emotion categories!
Psychological Models of Emotions
Psychological Theories of Basic Emotions

- Paul Ekman, 1971: Six Basic Emotions
- Plutchik, 1980: Eight Basic Emotions
- And many others
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*
Psychological Models of Emotions

- the valence, arousal, and dominance model
- the basic emotions model

We annotate data for both
We build automatic systems to detect both
Two Parts To The Work

Human annotations of words, phrases, tweets, etc. for emotions

- Draw inferences about people:
  - understand how we (or different groups of people) use language to express meaning and emotions

Develop automatic emotion related systems

- predicting emotions of words, tweets, sentences, etc.
- detecting stance, personality traits, well-being, cyber-bullying, etc.
Searching for Emotions (humans)
Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words
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:
Rating scales:

UNDERSTANDING ONLINE STAR RATINGS:

[HAS ONLY ONE REVIEW]

★★★★★
EXCELLENT

★★★★
OK

★★★*
CRAP

★★★

★★☆☆☆
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.

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

source: Wikimedia Commons
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)

\[
\text{score}(w) = \frac{\#\text{best}(w) - \#\text{worst}(w)}{\#\text{annotations}(w)}
\]

the scores range from:
-1 (least \(X\))  \(X = \text{property of interest, say valence}\)
to 1 (most \(X\))

- 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)
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)

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

Includes:

- Terms from the NRC Emotion Lexicon
- Terms from the Warriner et al. (2013) VAD lexicon
- Terms common in tweets
### Valence, Arousal, and Dominance Annotations (with BWS)

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

number of pairs of best—worst annotations
Best–Worst Scaling (Louviere & Woodworth, 1990)

- We can obtain real-valued scores for all the terms using a simple counting method (Orme, 2009)

\[
\text{score}(w) = \frac{\#\text{best}(w) - \#\text{worst}(w)}{\#\text{annotations}(w)}
\]

- 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
Entries with Highest and Lowest Scores in the VAD Lexicon

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

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)
Split-Half Reliability Scores for VAD Annotations

<table>
<thead>
<tr>
<th>Annotations</th>
<th># Terms</th>
<th># Annotations</th>
<th>V</th>
<th>A</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warriner et al. (2013)</td>
<td>13,915</td>
<td>20 per term</td>
<td>0.91</td>
<td>0.79</td>
<td>0.77</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Annotations</th>
<th># Terms</th>
<th># Annotations</th>
<th>V</th>
<th>A</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warriner et al. (2013)</td>
<td>13,915</td>
<td>20 per term</td>
<td>0.91</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>Ours (Warriner terms)</td>
<td>13,915</td>
<td>6 per tuple</td>
<td>0.95</td>
<td>0.91</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Split-Half Reliability Scores for VAD Annotations

<table>
<thead>
<tr>
<th>Annotations</th>
<th># Terms</th>
<th># Annotations</th>
<th>V</th>
<th>A</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warriner et al. (2013)</td>
<td>13,915</td>
<td>20 per term</td>
<td>0.91</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>Ours (Warriner terms)</td>
<td>13,915</td>
<td>6 per tuple</td>
<td>0.95</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Ours (all terms)</td>
<td>20,007</td>
<td>6 per tuple</td>
<td>0.95</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

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 $r$

<table>
<thead>
<tr>
<th>Annotations</th>
<th>V</th>
<th>A</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours-Warriner</td>
<td>0.81</td>
<td>0.62</td>
<td>0.33</td>
</tr>
<tr>
<td>(for overlapping terms)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The especially low correlations for dominance and arousal indicate that our lexicon has substantially different scores and rankings of terms.
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.
    - how different genders perceive and use language.
Analysis of VAD Judgments by Different Demographic Groups

Showed that our demographic attributes impact how we view the world around us. E.g.:

- women have a higher shared understanding of arousal of terms
- men have a higher shared understanding of dominance and valence
- those above the age of 35 have a higher shared understanding of V and A
- extroverts and those that are open to experiences have a higher shared understanding of V, A, and D

# Best-Worst Scaling Lexicons

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Language</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Affect Intensity Lexicon</td>
<td>English</td>
<td>General</td>
</tr>
<tr>
<td>2. SemEval-2015 English Twitter Sentiment Lexicon</td>
<td>English</td>
<td>Twitter</td>
</tr>
<tr>
<td>3. SemEval-2016 Arabic Twitter Sentiment Lexicon</td>
<td>Arabic</td>
<td>Twitter</td>
</tr>
<tr>
<td>4. Sentiment Composition Lexicon for Negators, Modals, and Adverbs (SCL-NMA)</td>
<td>English</td>
<td>General</td>
</tr>
<tr>
<td>5. Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP)</td>
<td>English</td>
<td>General</td>
</tr>
</tbody>
</table>

Lexicons and papers available at: [http://saifmohammad.com/WebPages/lexicons.html](http://saifmohammad.com/WebPages/lexicons.html)
English Twitter Lexicon: Examples sentiment scores obtained using BWS

<table>
<thead>
<tr>
<th>Term</th>
<th>Sentiment Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>awesomeness</td>
<td>0.827</td>
</tr>
<tr>
<td>#happygirl</td>
<td>0.625</td>
</tr>
<tr>
<td>cant waitttt</td>
<td>0.601</td>
</tr>
<tr>
<td>don't worry</td>
<td>0.152</td>
</tr>
<tr>
<td>not true</td>
<td>-0.226</td>
</tr>
<tr>
<td>cold</td>
<td>-0.450</td>
</tr>
<tr>
<td>#getagrip</td>
<td>-0.587</td>
</tr>
<tr>
<td>#sickening</td>
<td>-0.722</td>
</tr>
</tbody>
</table>

Term
-1 (most negative) to 1 (most positive)
NRC Emotion Lexicon

- Entries for 14,200
- Associations (0 or 1) with 8 basic emotions

Available at: www.saifmohammad.com

Paper:
Use of The NRC Emotion Lexicon

- For research by the scientific community
  - Computational linguistics, psychology, digital humanities, robotics, public health research, etc.

- To analyze text
  - Brexit tweets, Radiohead songs, Trump tweets, election debates,…
  - **Wishing Wall**, uses the NRC Emotion lexicon to visualize wishes.

  Displayed in:
  - Tekniska Museet, Stockholm, Sweden, 2014
  - Onassis Cultural Centre, Athens, Greece, 2015
  - Zorlu Centre, Istanbul, Turkey, 2016

- In commercial applications
Affect Intensity Lexicon:
About 6000 Words from the NRC Emotion Lexicon Annotated for Intensity of Emotion

Highest anger intensity entries:
- outraged 0.964
- brutality 0.959
- hatred 0.953

Lowest anger intensity entries:
- sisterhood 0.015
- musical 0.011
- tree 0.000

Highest fear intensity entries:
- torture 0.984
- terrorist 0.972
- horrific 0.969

Lowest fear intensity entries:
- volunteer 0.031
- lines 0.031
- romance 0.031

Scores are in the range 0 (lowest intensity) to 1 (highest intensity).

LREC-2018 Paper on the Relationship Between Basic Emotions and VAD

Dominance–Arousal scatter plots for words associated with the four emotions.


The size of the point is proportional to the intensity of the emotion.
Papers:

• **Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words.** Saif M. Mohammad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, July 2018.


• **The Effect of Negators, Modals, and Degree Adverbs on Sentiment Composition.** Svetlana Kiritchenko and Saif M. Mohammad, In *Proceedings of the NAACL 2016 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*, June 2014, San Diego, California.


@SaifMMohammad
Finding Emotions (machines)

- automatic systems for emotion, sentiment, personality, literary analysis, music generation,…
Visualizing Emotions in Text
Percentage of joy and anger words in close proximity to occurrences of man and woman in books.
Tracking Emotions in Stories

- Can we automatically track the emotions of characters?
- Are there some canonical shapes common to most stories?
- Can we track the change in distribution of emotion words?
Work on shapes of stories

- **From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales**, Saif Mohammad, In Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), June 2011, Portland, OR.


Generating music from text

Paper:

A method to generate music from literature.
• music that captures the change in the distribution of emotion words.
Music-Emotion Associations

- Major and Minor Keys
  - major keys: happiness
  - minor keys: sadness

- Tempo
  - fast tempo: happiness or excitement

- Melody
  - a sequence of consonant notes: joy and calm
  - a sequence of dissonant notes: excitement, anger, or unpleasantness

TransProse

Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.
TransProse

Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.

Examples
TransProse

Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.

Examples
TransProse

Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.

Examples

TransProse:  www.musicfromtext.com
Music played 300,000 times since website launched in April 2014.
TransProse Music Played by an Orchestra, at the Louvre Museum, Paris

A symphony orchestra performs under the glass of the Louvre museum in Paris on Sept. 20. Accenture Strategy has created a symphonic experience enabled by human insight and artificial intelligence technology. (Michel Euler/AP)
Debate: Universality of Perception of Emotions

Grad school experiment on people’s ability to distinguish photos of depression from anxiety
  ◦ one is based on sadness, and the other on fear
  ◦ found agreement to be poor

Margaret Mead
Cultural anthropologist

Paul Ekman
Psychologist and discoverer of micro expressions.

Lisa Barrett
University Distinguished Professor of Psychology, Northeastern University
Some Emotions more basic than others? may be not…
Hashtagged Tweets

- Hashtagged words are good labels of sentiments and emotions
  
  Some jerk just stole my photo on #tumblr #grrr #anger

- Hashtags are not always good labels:
  - hashtag used sarcastically
    
    The reviewers want me to re-annotate the data. #joy

Paper:

Data to Model Hundreds of Emotions

Papers:
- Using Nuances of Emotion to Identify Personality. Saif M. Mohammad and Svetlana Kiritchenko, In Proceedings of the ICWSM Workshop on Computational Personality Recognition, July 2013, Boston, USA.
Sentiment Lexicons

Created a sentiment lexicon using a Turney (2003) inspired method that uses PMI of a word with co-occurring positive and negative seed hashtags.

**Positive**
- spectacular 0.91
- okay 0.3

**Negative**
- lousy -0.74
- murder -0.95
SemEval Shared task on the Sentiment Analysis of Tweets

Papers:
• *NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets*, Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu, In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, June 2013, Atlanta, USA.
Sentiment Analysis Competition
SemEval-2013: Classify Tweets, 44 teams

F-score

Teams
Sentiment Analysis Competition
SemEval-2013: Classify SMS messages, 30 teams

F-score

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8

NRC-Canada  GUMLTLT  KLUE  AVAYA teragram  NTNU  CodeLab  SICT_EuALP  FBK-inst  ECHNUS  MTLDB  SAIL  UNITOR  SenseabilicTeam  NILC  USP  REACTION  SU-sentilab  LVIC-LMSI  FBM  OPTWIMA  senti-uen  ASUUnOLeipzig  SSA-UO  btkmuhlph  UMCC_DSL_(SA)  UoM  uottawa  IIRG
Feature Contributions (on Tweets)

F-scores

@SaifMMohammad
Detecting Stance in Tweets

Given a tweet text and a target determine whether:

- the tweeter is in favor of the given target
- the tweeter is against the given target
- neither inference is likely

Example 1:

Target: Donald Trump
Tweet: Jeb Bush is the only sane candidate in this republican lineup.

Systems have to deduce that the tweeter is likely against the target.

Example 2:

Target: pro-life movement
Tweet: The pregnant are more than walking incubators, and have rights!

Systems have to deduce that the tweeter is likely against the target.
SemEval-2018 Task 1: Affect in Tweets
https://competitions.codalab.org/competitions/17751

Tasks: Inferring likely affectual state of the tweeter
- emotion intensity regression (EI-reg)
- emotion intensity ordinal classification (EI-oc)
- sentiment intensity regression (V-reg)
- sentiment analysis, ordinal classification (V-oc)
- multi-label emotion classification task (E-c)

English, Arabic, and Spanish Tweets

75 Team (~200 participants)

### Participating Systems: ML algorithms

<table>
<thead>
<tr>
<th>ML algorithm</th>
<th>#Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EI-reg</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>1</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>10</td>
</tr>
<tr>
<td>CNN</td>
<td>10</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>8</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>11</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>9</td>
</tr>
<tr>
<td>LSTM</td>
<td>13</td>
</tr>
<tr>
<td>Random Forest</td>
<td>8</td>
</tr>
<tr>
<td>RNN</td>
<td>0</td>
</tr>
<tr>
<td>SVM or SVR</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>14</td>
</tr>
</tbody>
</table>
## Participating Systems: features

<table>
<thead>
<tr>
<th>Features/Resources</th>
<th>#Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EI-reg</td>
</tr>
<tr>
<td>affect-specific word embeddings</td>
<td>10</td>
</tr>
<tr>
<td>affect/sentiment lexicons</td>
<td>24</td>
</tr>
<tr>
<td>character ngrams</td>
<td>6</td>
</tr>
<tr>
<td>dependency/parse features</td>
<td>2</td>
</tr>
<tr>
<td>distant-supervision corpora</td>
<td>10</td>
</tr>
<tr>
<td>manually labeled corpora (other)</td>
<td>6</td>
</tr>
<tr>
<td>AIT-2018 train-dev (other task)</td>
<td>6</td>
</tr>
<tr>
<td>sentence embeddings</td>
<td>10</td>
</tr>
<tr>
<td>unlabeled corpora</td>
<td>6</td>
</tr>
<tr>
<td>word embeddings</td>
<td>32</td>
</tr>
<tr>
<td>word ngrams</td>
<td>19</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
</tbody>
</table>
SemEval-2018 Task 1: Affect in Tweets
https://competitions.codalab.org/competitions/17751

Tasks: Inferring likely affectual state of the tweeter
- emotion intensity regression
- emotion intensity ordinal classification
- sentiment intensity regression
- sentiment analysis, ordinal classification
- emotion classification task

English, Arabic, and Spanish Tweets

75 Team (~200 participants)

Includes a separate evaluation component for biases towards race and gender.
Do Machines Make Fair Decisions?

YES:
- they do not take bribes
- they can make decisions without being influenced by the user's gender, race, or sexual orientation

And NO—recent studies have demonstrated that predictive models built on historical data may inadvertently inherit inappropriate human biases
Do Machines Make Fair Decisions?

YES:

• they do not take bribes
• they can make decisions without being influenced by the user's gender, race, or sexual orientation

And NO—recent studies have demonstrated that predictive models built on historical data may inadvertently inherit inappropriate human biases
Occurrences of “son” and “daughter” in the Google Books Ngram corpus
Occurrences of “genius son” and “genius daughter” in the Google Books Ngram corpus
Showed that parents search disproportionately more on Google for:

- is my son gifted? than is my daughter gifted?
- is my daughter overweight? than is my son overweight?
Previous Studies

- focus on one or two systems or resources
  - word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Speer, 2017)
- no benchmark dataset for examining inappropriate biases

Our Work

- **Equity Evaluation Corpus (EEC)**—a dataset of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders
- using the EEC, examine the output of 219 sentiment analysis systems that took part in the SemEval-2018 Affect in Tweets shared task
**Race Bias Results:** Box plot of the score differences on the AA-EA name sentence pairs for each system on the valence regression task (plots for the four emotions are similar).

Bias Results

- more than 75% of the systems tend to consistently mark sentences involving one gender/race with higher intensity scores
- biases are more common for race than for gender
- bias can be different depending on the affect dimension involved
- score differences are small on average (about 3% of the 0 to 1 score range)
- for some systems the score differences reached as high as 34% of the range
- score differences may be higher for complex sentences involving many gender-/race-associated words

Art and Emotions

Art and Emotions

- Art is imaginative human creation meant to evoke an emotional response
- Large amounts of art are now online
  - With title, painter, style, year, etc.
  - Not labeled for emotions evoked
- Useful:
  - Ability to search for paintings evoking the desired emotional response
  - Automatically detect emotions evoked by paintings
  - Automatically transform (or generate new) paintings
  - Identify what makes paintings evocative

Figure 1: WikiArt.org’s page for the Mona Lisa. In the WikiArt Emotions Dataset, the Mona Lisa is labeled as evoking happiness, love, and trust; its average rating is 2.7 (in the range of –3 to 3).
WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art

- ~4K pieces of art (mostly paintings)
- From four styles: Renaissance Art, Post-Renaissance Art, Modern Art, and Contemporary Art
- 20 categories: Impressionism, Expressionism, Cubism, Figurative art, Realism, Baroque,…
- Annotated for emotions evoked, amount liked, does it depict a face.

Figure 1: WikiArt.org’s page for the Mona Lisa. In the WikiArt Emotions Dataset, the Mona Lisa is labeled as evoking happiness, love, and trust; its average rating is 2.1 (in the range of −3 to 3).

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.
Summary

- Created several lexicons that capture word-emotion associations
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

The NRC Emotion Intensity Lexicon aka Affect Intensity Lexicon
provides intensity scores for ~6000 words with four emotions
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
Summary

- Created several lexicons that capture word-emotion associations
- Used best-worst scaling to obtain reliable real-valued scores
- Showed that the lexicons have a wide variety of creative applications and can be used to understand creativity in language
- Showed that different demographic groups may have different shared understandings of the world around us
- Proposed new tasks such as stance detection and emotion intensity detection
- Investigated aspects of fairness and bias in language

As creators of NLP/ML systems, we have the opportunity and responsibility to:
- measure and address aspects of bias and fairness
- better understand human creativity
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
- Sentiment and emotion lexicons and corpora
- Links to shared tasks
- Interactive visualizations
- Tutorials and book chapters on sentiment and emotion analysis

Saif M. Mohammad
✉ Saif.Mohammad@nrc-cnrc.gc.ca
🐦 @SaifMMohammad