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
THE ORIGIN OF SPECIES
BY MEANS OF NATURAL SELECTION,
OR THE
PRESERVATION OF FAVOURED RACES IN THE STRUGGLE
FOR LIFE

By CHARLES DARWIN, M.A.
Psychological Theories of Basic Emotions

- Paul Ekman, 1971: Six Basic Emotions
- Plutchik, 1980: Eight Basic Emotions
- And many others

Plutchik’s Emotion Wheel
Image credit: Julia Belyanevych

Paul Ekman, Psychologist
Core Dimensions of Connotative 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 work with both models
Two Parts To The Work

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

- Draw inferences about language and 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.
The Search for Emotions – by humans
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
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
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 scales
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 scales
Rating scales:

1. It might be an itch
2. I just need a bandaid
3. It's kind of annoying
4. This is concerning but I can still work
5. Bees?
6. Bees!
7. I can't stop crying
8. I can't move it hurts so bad
9. Mauled by a bear or ninjas
10. Unconscious

source: imgur
Rating scales:

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.
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)*

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

the scores range from:

- 1 (least X)  
  - X = property of interest, say valence
- 1 (most X)  

- the scores can then be used to rank all the terms
Best–Worst Scaling (Louviere & Woodworth, 1990)

- Uses comparative annotation—mitigates bias
- 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 Questionnaire for Valence Annotations

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><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td>40,014</td>
<td></td>
<td></td>
<td></td>
<td><strong>778,085</strong></td>
</tr>
</tbody>
</table>

Includes:

- Terms from the NRC Emotion Lexicon
- Terms from the Warriner et al. (2013) VAD lexicon
- Terms common in tweets
Best–Worst Scaling (Louviere & Woodworth, 1990)

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

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

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

  - linearly transformed to scores between 0 and 1
  - 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)

Pearson correlation: -1 (most inversely correlated) to 1 (most correlated), higher scores indicate higher reliability
Split-Half Reliability Scores for VAD Annotations

higher scores indicate higher reliability

<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

Higher scores indicate higher reliability.

<table>
<thead>
<tr>
<th>Annotations</th>
<th># Terms</th>
<th># Annot.</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>
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 (for overlapping terms)</td>
<td>0.81</td>
<td>0.62</td>
<td>0.33</td>
</tr>
</tbody>
</table>

The especially low correlations for dominance and arousal indicate that our lexicon has substantially different scores and rankings of terms.
Gender Differences in Perception of the World (Core Dimensions)

- 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

About 6000 Words from the NRC Emotion Lexicon Annotated for Intensity of Emotion

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

The Search for Emotions – by Machines

- automatic systems for detecting emotions in text, literary analysis, music generation, …
Detecting Emotions in Stories
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?
As You Like It

Hamlet

Frankenstein
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
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.

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

<table>
<thead>
<tr>
<th>Teams</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRC-Canada</td>
<td>0.8</td>
</tr>
<tr>
<td>OUMU-LTL</td>
<td>0.7</td>
</tr>
<tr>
<td>teragram</td>
<td>0.7</td>
</tr>
<tr>
<td>AIVAYA</td>
<td>0.6</td>
</tr>
<tr>
<td>BOUNCE</td>
<td>0.6</td>
</tr>
<tr>
<td>KLLE</td>
<td>0.6</td>
</tr>
<tr>
<td>AMI_and_FRC</td>
<td>0.6</td>
</tr>
<tr>
<td>FBM</td>
<td>0.6</td>
</tr>
<tr>
<td>SALM</td>
<td>0.6</td>
</tr>
<tr>
<td>UTD-DB</td>
<td>0.6</td>
</tr>
<tr>
<td>PBK-LF</td>
<td>0.6</td>
</tr>
<tr>
<td>UNITOR</td>
<td>0.6</td>
</tr>
<tr>
<td>ECNOC-UofB</td>
<td>0.6</td>
</tr>
<tr>
<td>LVIC-LIMS</td>
<td>0.6</td>
</tr>
<tr>
<td>NILC-USP</td>
<td>0.6</td>
</tr>
<tr>
<td>DariMinning</td>
<td>0.6</td>
</tr>
<tr>
<td>UT-DB</td>
<td>0.6</td>
</tr>
<tr>
<td>ASUUnioUlaipioz2010</td>
<td>0.6</td>
</tr>
<tr>
<td>OPT-MMA</td>
<td>0.6</td>
</tr>
<tr>
<td>bwarleigh</td>
<td>0.6</td>
</tr>
<tr>
<td>SZE-ANLP</td>
<td>0.6</td>
</tr>
<tr>
<td>codeX</td>
<td>0.6</td>
</tr>
<tr>
<td>Oasis</td>
<td>0.6</td>
</tr>
<tr>
<td>NTNU</td>
<td>0.6</td>
</tr>
<tr>
<td>UoM</td>
<td>0.6</td>
</tr>
<tr>
<td>SSA-UO</td>
<td>0.6</td>
</tr>
<tr>
<td>SenemylcTeam</td>
<td>0.6</td>
</tr>
<tr>
<td>UMCC_DLSI_SAsaamu</td>
<td>0.6</td>
</tr>
<tr>
<td>shinai</td>
<td>0.6</td>
</tr>
<tr>
<td>sentlass-en</td>
<td>0.6</td>
</tr>
<tr>
<td>SU-sentilab</td>
<td>0.6</td>
</tr>
<tr>
<td>REACTION</td>
<td>0.6</td>
</tr>
<tr>
<td>uottawa</td>
<td>0.6</td>
</tr>
<tr>
<td>IIT-B-SentiAnalyses</td>
<td>0.6</td>
</tr>
<tr>
<td>IIRG</td>
<td>0.6</td>
</tr>
</tbody>
</table>

@SaifMMohammad
Sentiment Analysis Competition
SemEval-2013: Classify SMS messages, 30 teams
Feature Contributions (on Tweets)

F-scores

@SaifMMohammad
The Sentiment Analysis that Companies Want!
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 viable candidate in this republican lineup.
Systems have to deduce that the tweeter is likely against the target.

Example 2:
Target: pro-choice movement
Tweet: The pregnant are more than walking incubators, and have rights!
Systems have to deduce that the tweeter is likely in favour 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>El-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>EL-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 as the models have become more sophisticated, they have inadvertently inherited inappropriate human biases.
Examples of Biased AI

- Tay, Microsoft’s racist chat bot posting inflammatory and offensive tweets
- Amazon’s AI recruiting tool biased against women
  - penalized resumes that included the word “women’s,” as in “women’s chess club captain”
- Face recognition systems good for detecting faces of white men, but really bad for African American women
- Recidivism systems that are biased against people from African American neighborhoods

@SaifMMohammad built on human data
Examples of Biased AI

- Tay, Microsoft’s racist chat bot posting inflammatory and offensive tweets
- Amazon’s AI recruiting tool biased against women
  - penalized resumes that included the word “women’s,” as in “women’s chess club captain”
- Face recognition systems good for detecting faces of white men, but really bad for African American women
- Recidivism systems that are biased against people from African American neighborhoods
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?
Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems

- Found that most systems consistently give higher emotion intensity scores to sentences when they have mentions of one race/gender as opposed to another race/gender

We Need More Work On

- Measuring inappropriate biases in AI systems and inappropriate biases in language
- Mitigating inappropriate biases in automatic systems
Ongoing Work

- The Natural Selection of Words
  - How words compete to represent a meaning

Art and Emotions

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.

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.
The State of NLP Literature: A Diachronic Analysis of the ACL Anthology

- Part I: Size and demographics
- Part II: Areas of Research (Examining Title Terms)
- Part III: Impact — Overall, and across Paper Types and Venues
The State of NLP Literature: A Diachronic Analysis of the ACL Anthology
The State of NLP Literature: A Diachronic Analysis of the ACL Anthology
Percentage of Female First Author (FFA) Papers

Note: When the number of papers is small (e.g., between 1965 and 1980 or when selecting papers that have a less common title word), then the percentages are not very meaningful and those areas of the graph can show large variations.

Common Title words and their FFA% (Select any of the words below or enter word in the text box in the top right to view FFA% for those papers.)

Title words in order of FFA%

<table>
<thead>
<tr>
<th>Title word</th>
<th>FFA%</th>
</tr>
</thead>
<tbody>
<tr>
<td>discourse</td>
<td>49.9%</td>
</tr>
<tr>
<td>annotation</td>
<td>42.6%</td>
</tr>
<tr>
<td>study</td>
<td>40.9%</td>
</tr>
<tr>
<td>corpus</td>
<td>40.0%</td>
</tr>
<tr>
<td>evaluation</td>
<td>39.4%</td>
</tr>
<tr>
<td>dialogue</td>
<td>38.9%</td>
</tr>
<tr>
<td>domain</td>
<td>38.7%</td>
</tr>
<tr>
<td>resources</td>
<td>37.6%</td>
</tr>
<tr>
<td>automatic</td>
<td>37.2%</td>
</tr>
<tr>
<td>spoken</td>
<td>37.1%</td>
</tr>
<tr>
<td>English</td>
<td>37.0%</td>
</tr>
<tr>
<td>linguistic</td>
<td>36.9%</td>
</tr>
<tr>
<td>identification</td>
<td>34.0%</td>
</tr>
<tr>
<td>systems</td>
<td>33.2%</td>
</tr>
<tr>
<td>data</td>
<td>32.9%</td>
</tr>
<tr>
<td>approach</td>
<td>32.7%</td>
</tr>
<tr>
<td>corpus</td>
<td>32.6%</td>
</tr>
<tr>
<td>detection</td>
<td>32.2%</td>
</tr>
<tr>
<td>cross</td>
<td>32.2%</td>
</tr>
<tr>
<td>information</td>
<td>32.0%</td>
</tr>
<tr>
<td>analysis</td>
<td>31.8%</td>
</tr>
<tr>
<td>syntactic</td>
<td>31.4%</td>
</tr>
</tbody>
</table>

FFA% in papers that have a given word in the title

<table>
<thead>
<tr>
<th>Word</th>
<th>FFA%</th>
</tr>
</thead>
<tbody>
<tr>
<td>analysis</td>
<td>62.6%</td>
</tr>
<tr>
<td>cross</td>
<td>58.2%</td>
</tr>
<tr>
<td>English</td>
<td>57.0%</td>
</tr>
<tr>
<td>inform</td>
<td>55.0%</td>
</tr>
<tr>
<td>learn</td>
<td>54.0%</td>
</tr>
<tr>
<td>machine</td>
<td>52.9%</td>
</tr>
<tr>
<td>annotation</td>
<td>32.9%</td>
</tr>
<tr>
<td>data</td>
<td>32.9%</td>
</tr>
<tr>
<td>entity</td>
<td>32.9%</td>
</tr>
<tr>
<td>model</td>
<td>32.9%</td>
</tr>
<tr>
<td>neural</td>
<td>32.9%</td>
</tr>
<tr>
<td>natural</td>
<td>32.9%</td>
</tr>
<tr>
<td>sematic</td>
<td>31.1%</td>
</tr>
<tr>
<td>syntactic</td>
<td>31.1%</td>
</tr>
<tr>
<td>unsupervised</td>
<td>31.1%</td>
</tr>
</tbody>
</table>

Content is visualized using a graph showing the percentage of FFA papers over the years and the percentage by title word. The graph includes a table listing common title words and their corresponding FFA% values.
The State of NLP Literature: Part I

Size and Demographics

Saif M. Mohammad
Oct 22 · 16 min read

This series of posts presents a diachronic analysis of the ACL Anthology — Or, as I like to think of it, making sense of NLP Literature through pictures.

The world of scientific publishing is a rain forest: Where ideas compete for sunlight/attention; Where some win out and grow taller, while others are forgotten. (Photo credit: Héctor J. Rivas)
The State of NLP Literature: This series of posts presents a diachronic analysis of the ACL Anthology — Or, as I like to think of it, making sense of NLP Literature through pictures. #NLProc #nlpScholar #inclusiveness

https://medium.com/@nlpScholar/state-of-nlp-cbf768492f90 ...

pic.twitter.com/AbauYhUibE

Impressions: 20,040
Times people saw this Tweet on Twitter

Total engagements: 1,770
Times people interacted with this Tweet

Reach a bigger audience
Get more engagements by promoting this Tweet!
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
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)

The NRC Emotion Intensity Lexicon aka Affect Intensity Lexicon
provides intensity scores for ~6000 words with four emotions
http://saifmohammad.com/WebPages/AffectIntensity.htm
(anger, fear, joy, sadness)

The NRC Word–Colour Association Lexicon
provides associations for ~14,000 words with 11 common colours
http://saifmohammad.com/WebPages/lexicons.html