The Search for Emotions, Creativity, and Fairness in Language

Saif M. Mohammad

Senior Research Scientist, National Research Council Canada

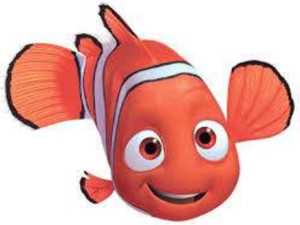
✉ Saif.Mohammad@nrc-cnrc.gc.ca @SaifMMohammad

National Research Council Canada
Conseil national de recherches Canada

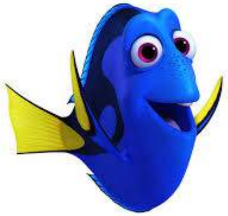
Canada

Emotions

- Determine human experience and behavior
- Condition our actions
- Central in organizing meaning
    - No cognition without emotion



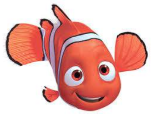
The Search for Emotions in Language



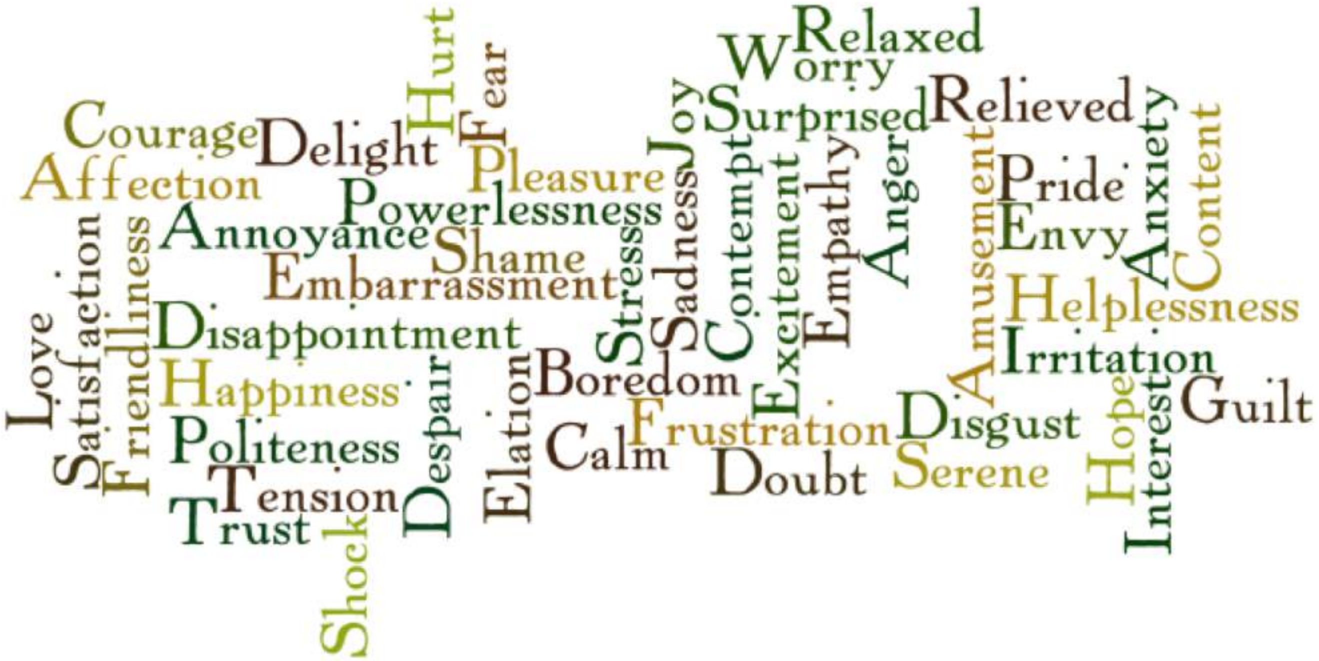
creativity

fairness

Emotions



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

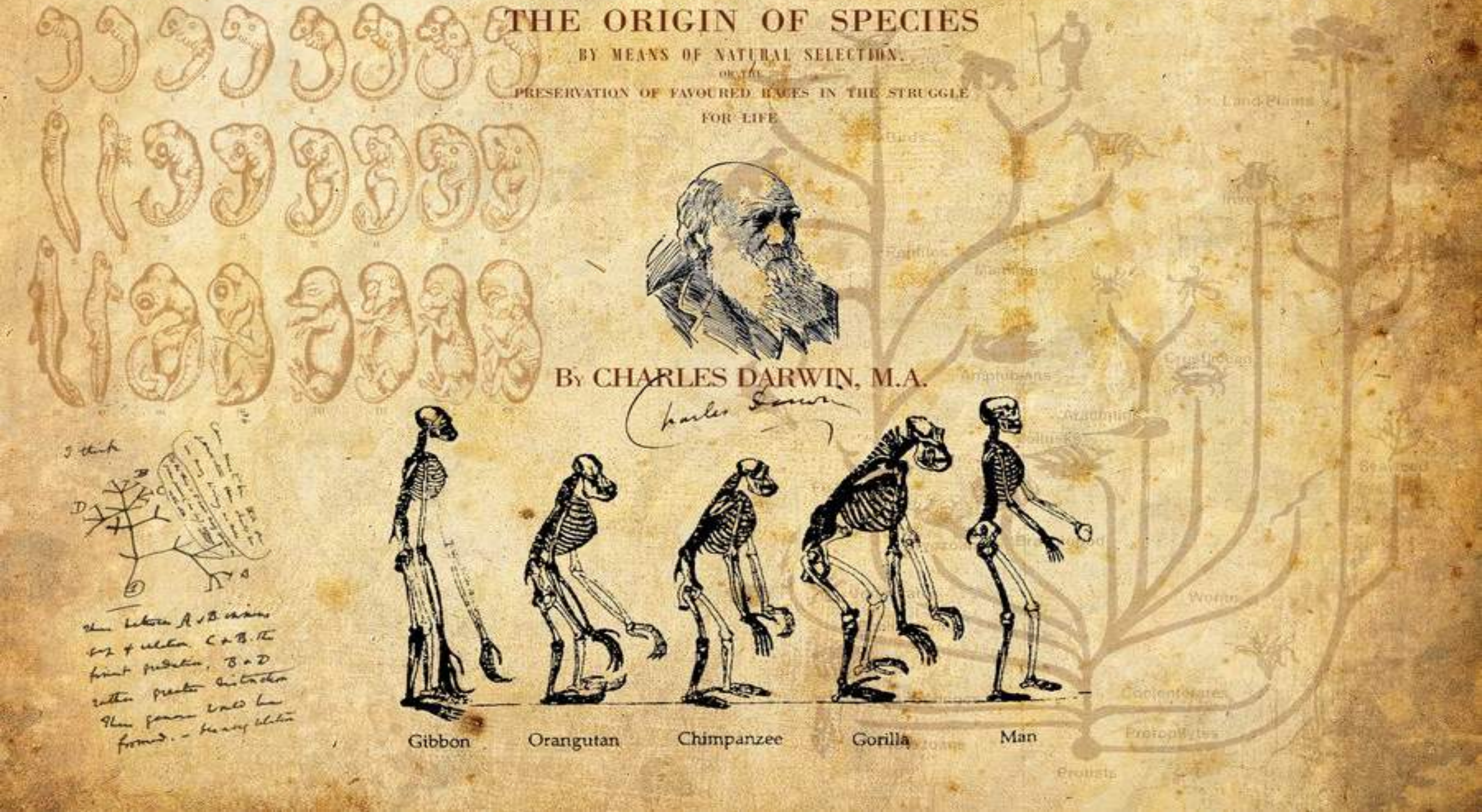
ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

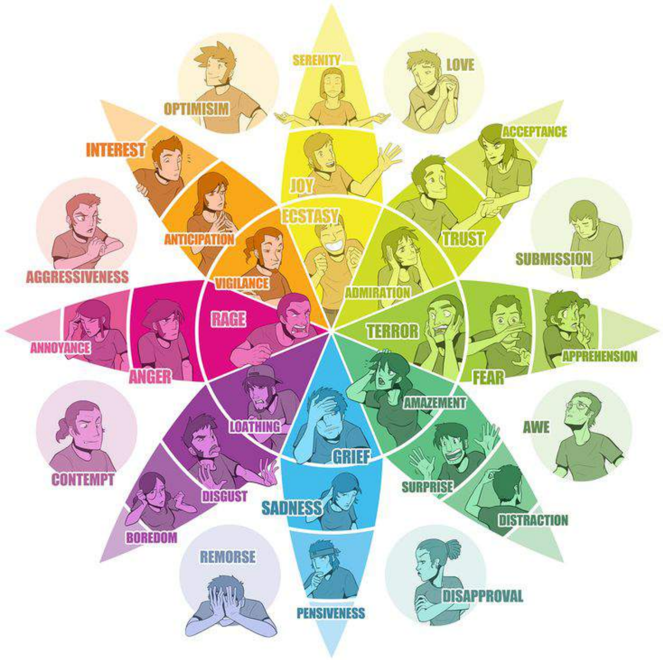
OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE



Psychological Theories of Basic Emotions

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

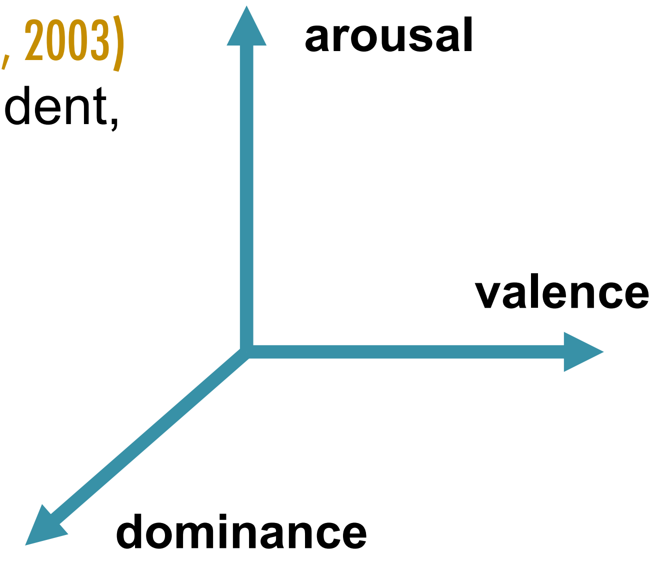


Plutchik's Emotion Wheel
Image credit: Julia Belyanevych

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

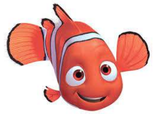
We annotate data for both

We build automatic systems to detect both

Two Parts To The Work

The Search for Emotions – by Humans

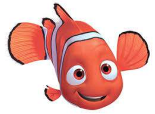
Human annotations of words, phrases, tweets, etc. for emotions



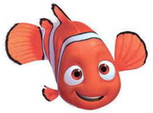
- Draw inferences about language and people:
  - understand how we (or different groups of people) use language to express meaning and emotions

The Search for Emotions – by Machines

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

Peter Turney

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

Available at: www.saifmohammad.com

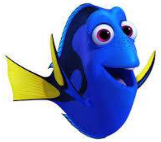
Paper:

Crowdsourcing a Word-Emotion Association Lexicon, Saif M. Mohammad and Peter Turney, *Computational Intelligence*, 29 (3), pages 436-465, 2013. Lexicon Released in 2010.

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:
    - **Barbican Centre**, London, England, 2014
    - **Tekniska Museet**, Stockholm, Sweden, 2014
    - **Onassis Cultural Centre**, Athens, Greece, 2015
    - **Zorlu Centre**, Istanbul, Turkey, 2016



creativity

- In commercial applications

fine-grained

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 scales

Related Work: Existing VAD Lexicons

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

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

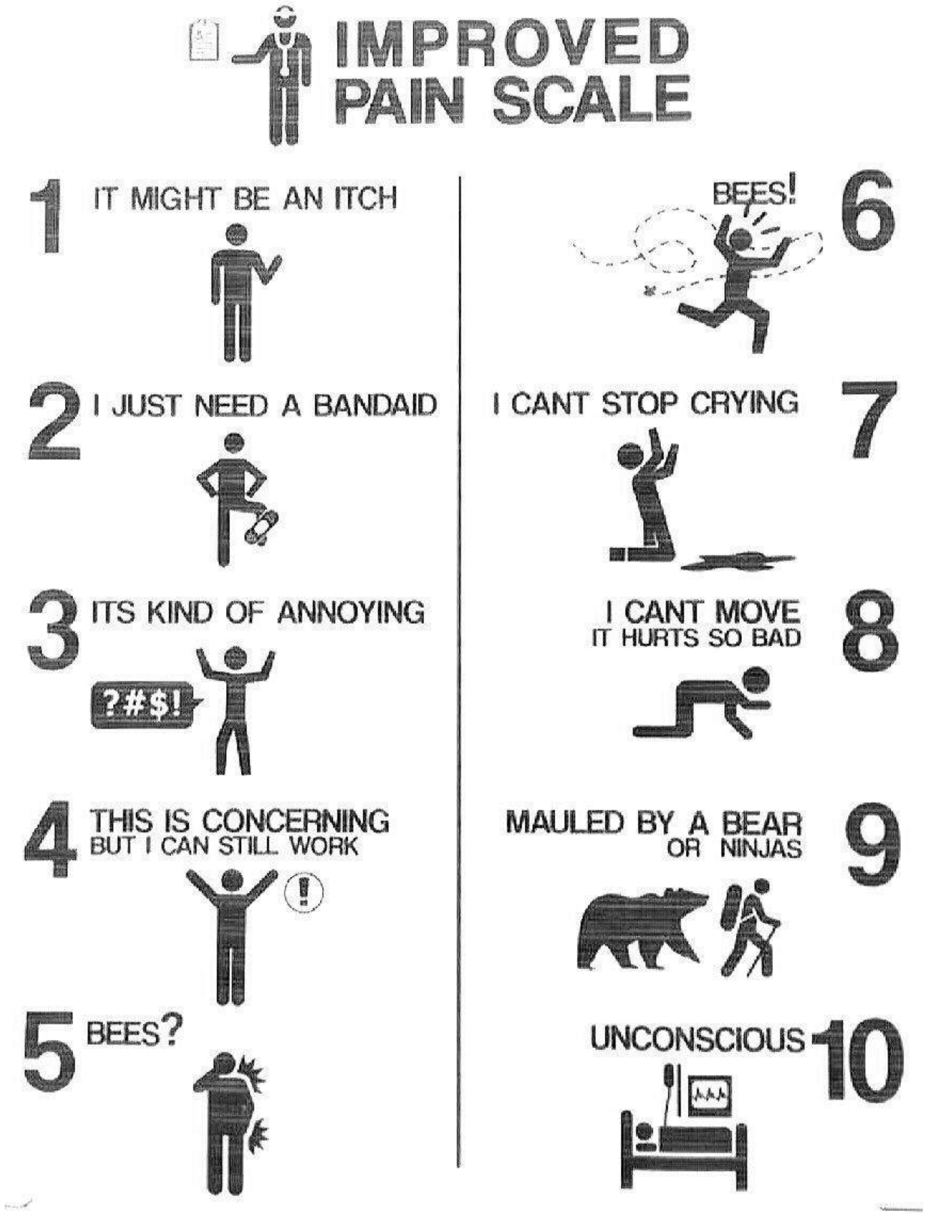
Warriner et al. Norms (Warriner et al. 2013)

- 14,000 words
- 9-point rating scale

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

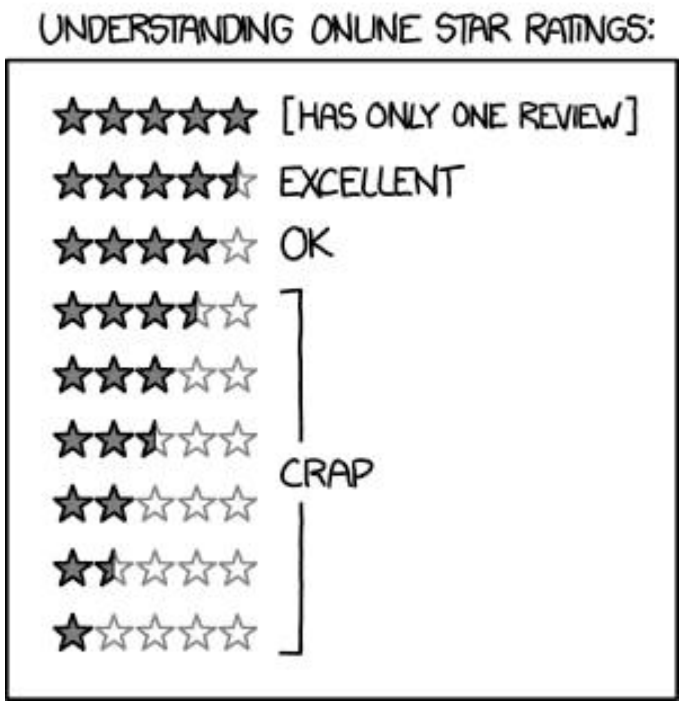
- E.g.:
  - Moors et al. (2013) for Dutch
  - Vo et al. (2009) for German
  - Redondo et al. (2007) for Spanish
- rating scales

Rating scales:



source: imgur

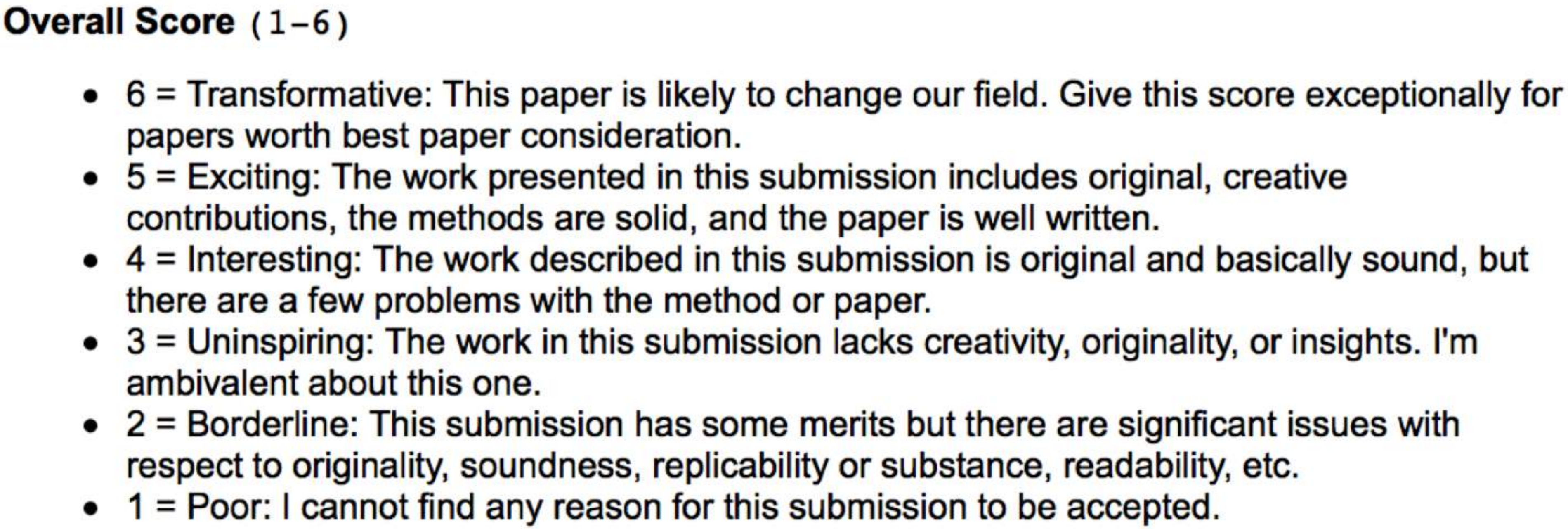
Rating scales:


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

source: xkcd

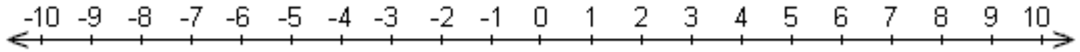
Rating scales:

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.



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)   X = 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)

- 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 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.

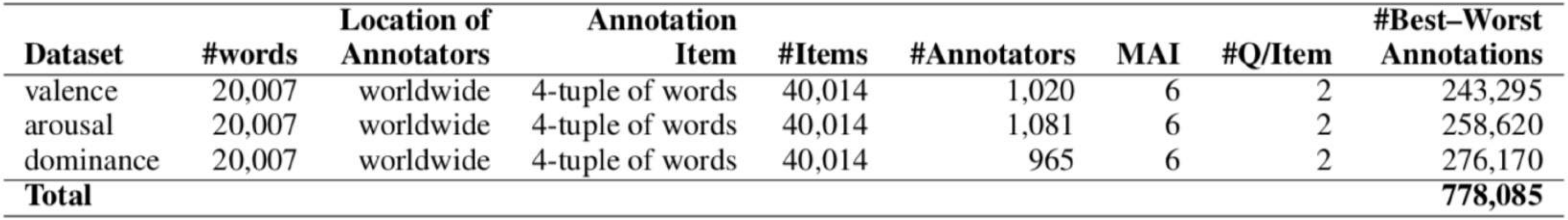
Crowdsourcing and Quality Control

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)

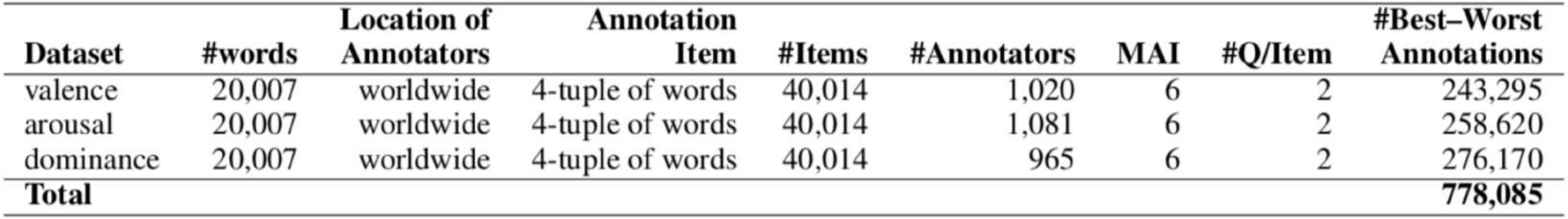


Dataset	#words	Location of Annotators	Annotation Item	#Items	#Annotators	MAI	#Q/Item	#Best–Worst Annotations
valence	20,007	worldwide	4-tuple of words	40,014	1,020	6	2	243,295
arousal	20,007	worldwide	4-tuple of words	40,014	1,081	6	2	258,620
dominance	20,007	worldwide	4-tuple of words	40,014	965	6	2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

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)



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)

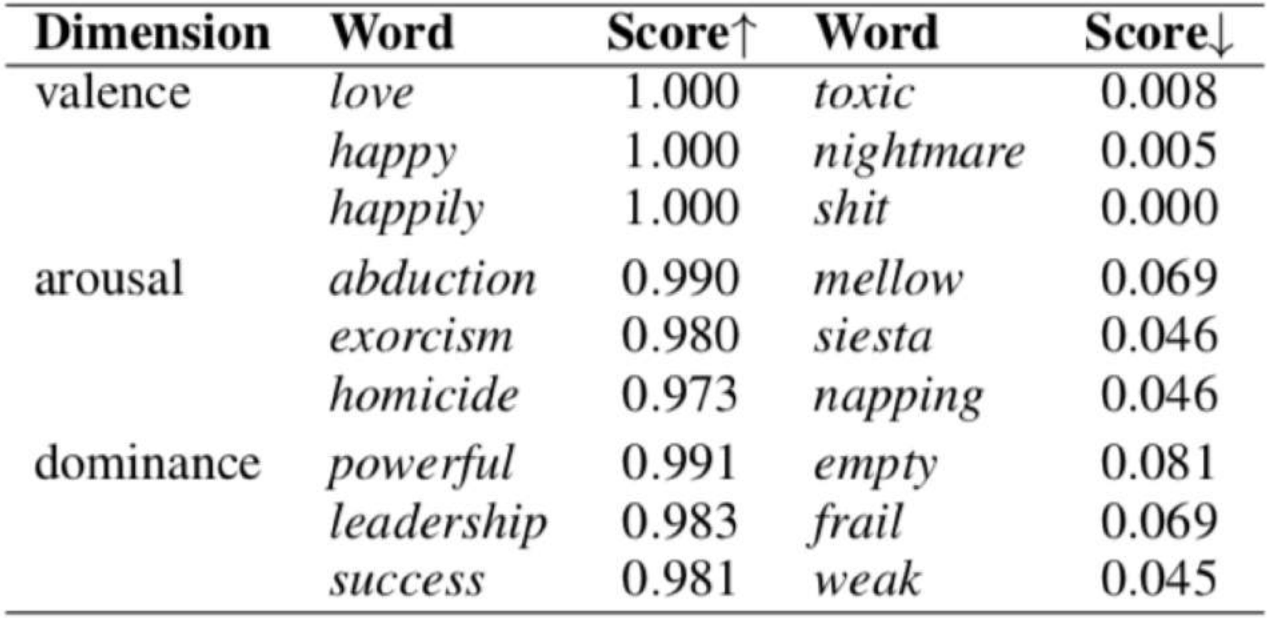
- We can obtain real-valued scores for all the terms using a simple counting method (Orme, 2009)

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

  the scores range from:
  -1 (least X)   X = property of interest, say valence
  to 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

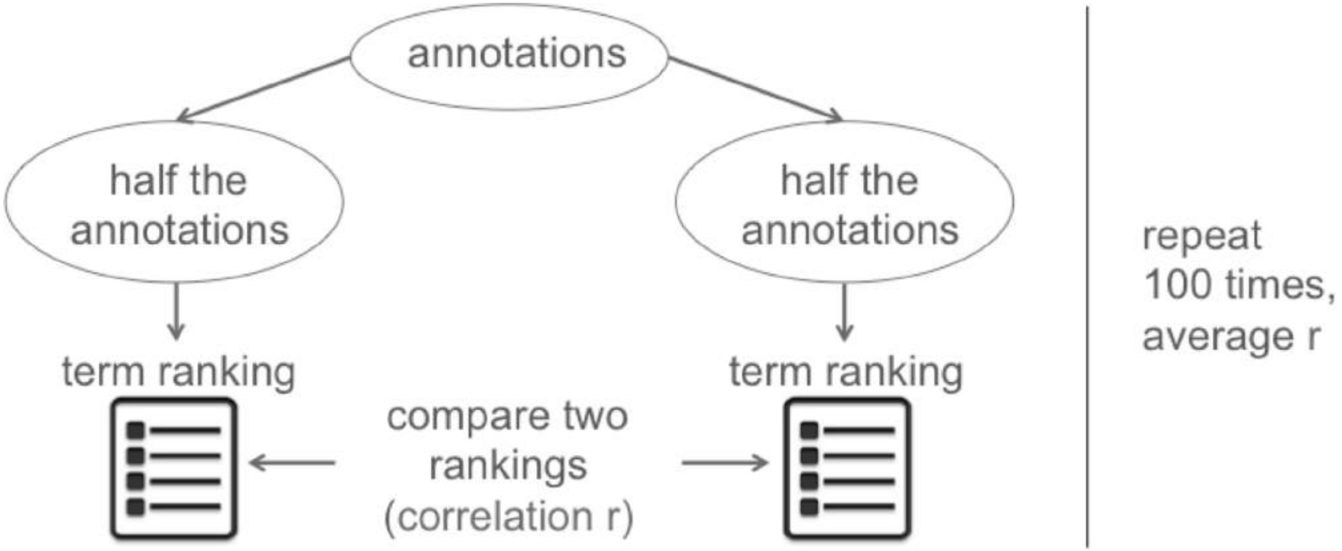


Dimension	Word	Score↑	Word	Score↓
valence	*love*	1.000	*toxic*	0.008
	happy	1.000	*nightmare*	0.005
	happily	1.000	*shit*	0.000
arousal	*abduction*	0.990	*mellow*	0.069
	exorcism	0.980	*siesta*	0.046
	homicide	0.973	*napping*	0.046
dominance	*powerful*	0.991	*empty*	0.081
	leadership	0.983	*frail*	0.069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

Split-Half Reliability Scores for VAD Annotations

higher scores indicate higher reliability

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

higher scores indicate higher reliability

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

higher scores indicate higher reliability

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.

Shared Understanding of VAD:
Within and Across Demographic Groups

fairness

- 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

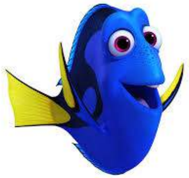
Analysis of VAD Judgments by Different Demographic Groups

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

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.

Best-Worst Scaling Lexicons



creativity

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

Lexicons and papers available at:
http://saifmohammad.com/WebPages/lexicons.html

Affect/Emotion 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

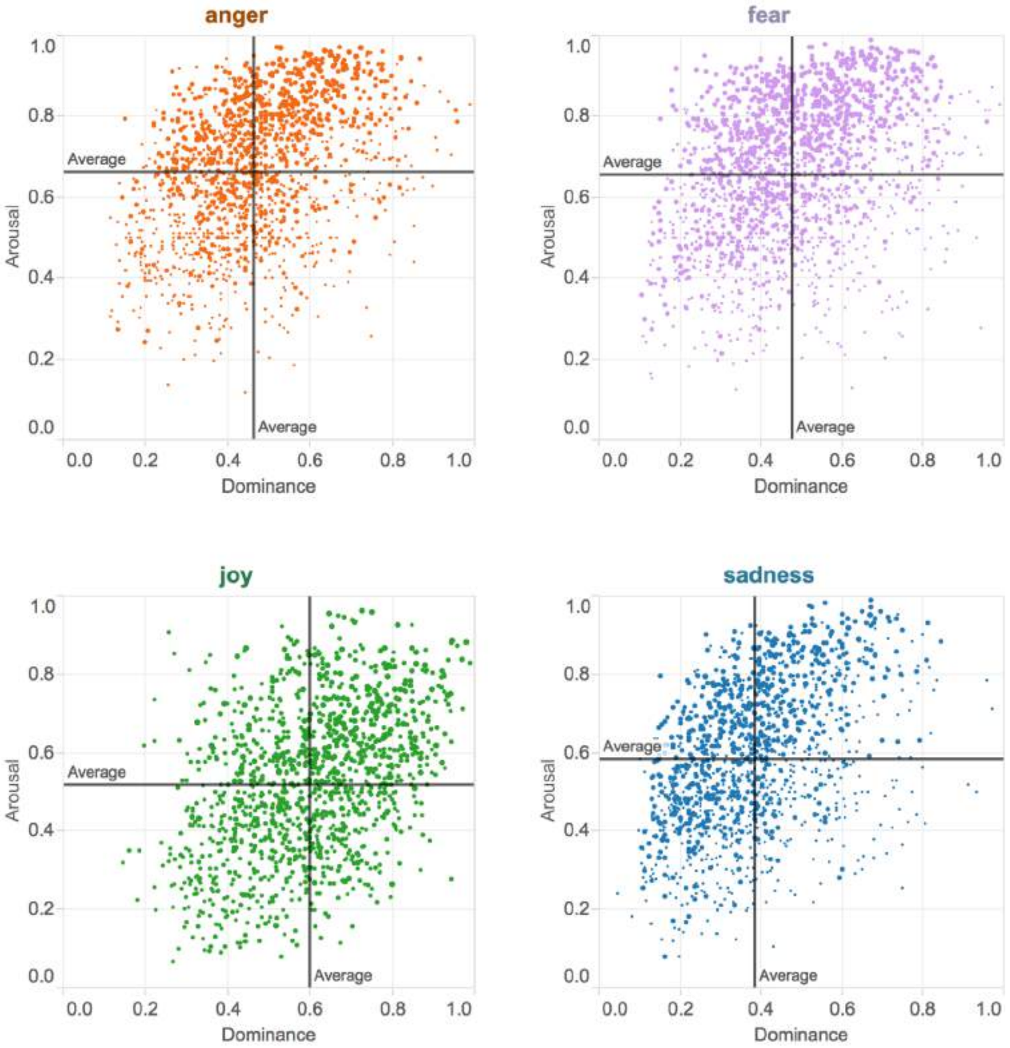
Scores are in the range 0 (lowest intensity) to 1 (highest intensity).

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

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

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

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



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

English Twitter Lexicon:
Examples sentiment scores obtained using BWS

Term	Sentiment Score -1 (most negative) to 1 (most positive)
awesomeness	0.827
#happygirl	0.625
cant waitttt	0.601
don't worry	0.152
not true	-0.226
cold	-0.450
#getagrip	-0.587
#sickening	-0.722

Papers:

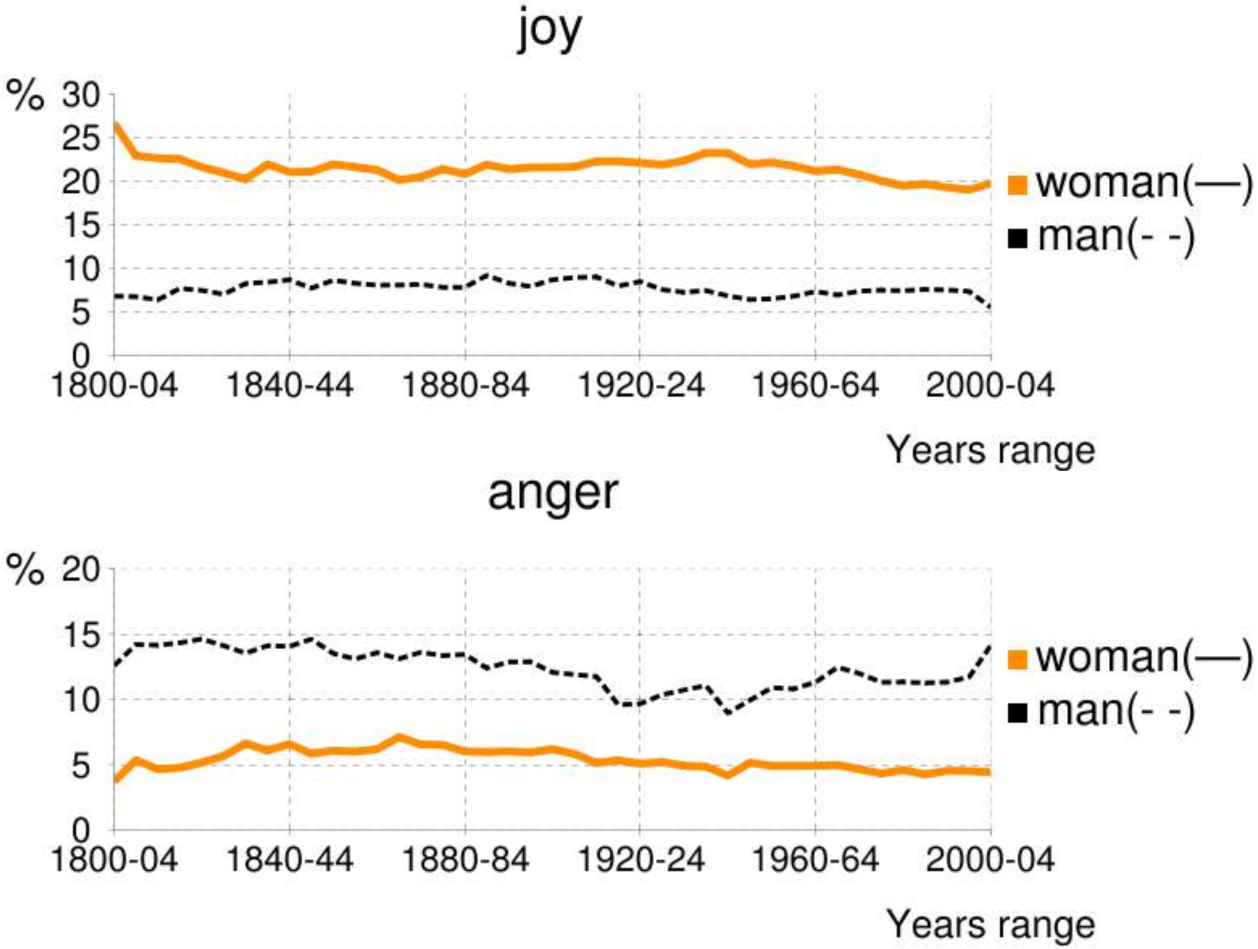
- Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Saif M. Mohammad. In *Proceedings of* the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, July 2018.
- Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling. Svetlana Kiritchenko and Saif M. Mohammad. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. June 2016. San Diego, CA.
- Word Affect Intensities. Saif M. Mohammad. In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), May 2018, Miyazaki, Japan.
- Sentiment Composition of Words with Opposing Polarities. Svetlana Kiritchenko and Saif M. Mohammad. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. June 2016. San Diego, CA.
- 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.
- Semeval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases. Svetlana Kiritchenko, Saif M. Mohammad, and Mohammad Salameh. In Proceedings of the International Workshop on Semantic Evaluation (SemEval '16). June 2016. San Diego, California.

The Search for Emotions – by Machines

- automatic systems for emotion, sentiment, personality, literary analysis, music generation,…

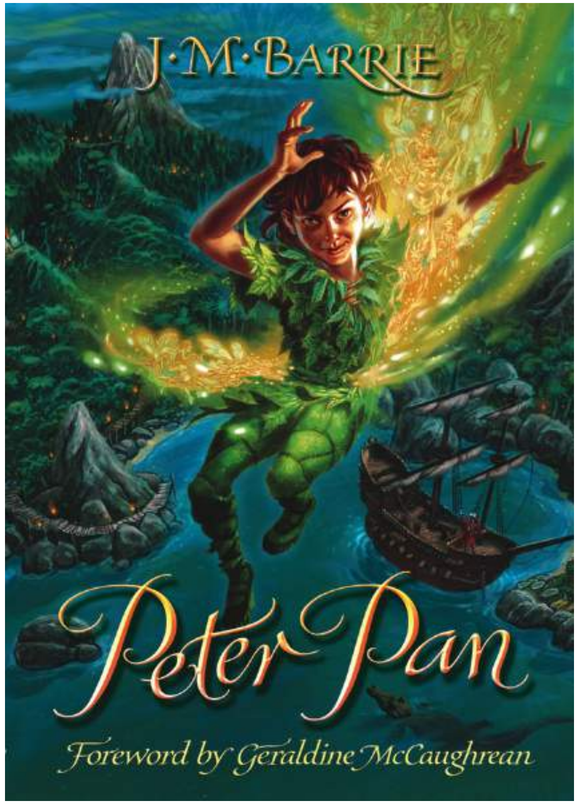
Tony Yang, Simon Fraser University

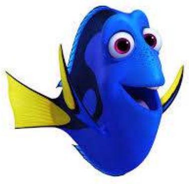
Visualizing Emotions in Text



fairness

Percentage of joy and anger words in close proximity to occurrences of man and woman in books.





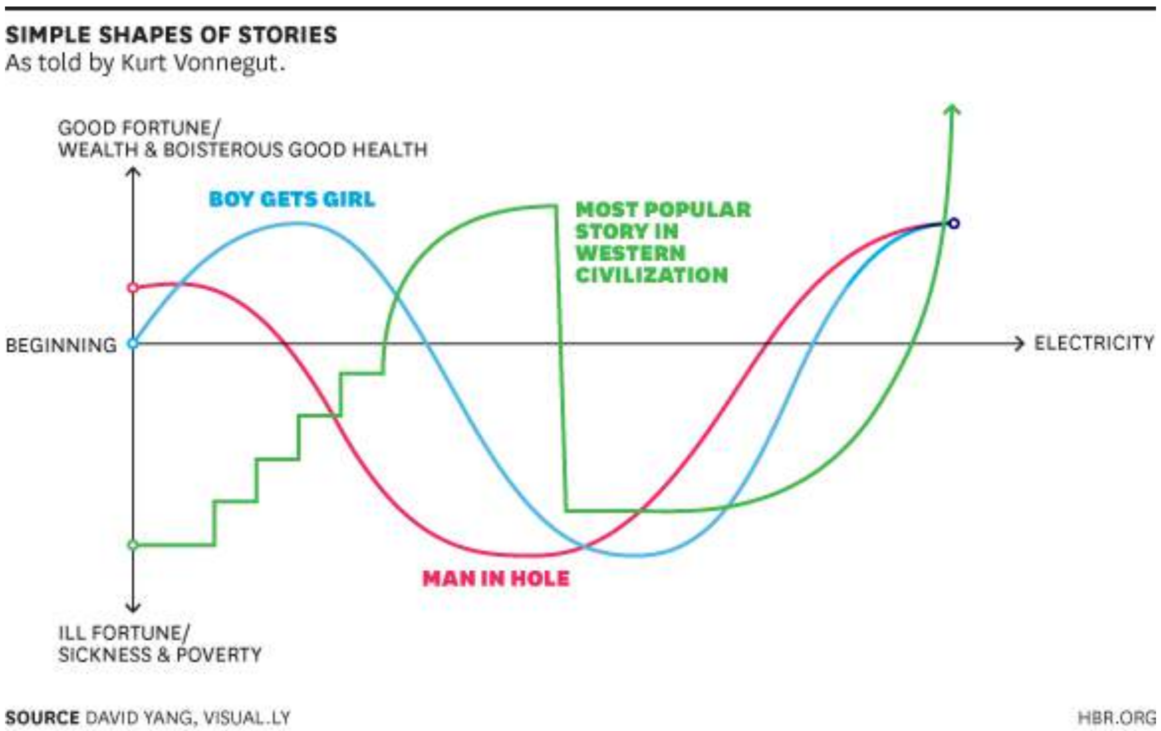
creativity

Stories

STORIES

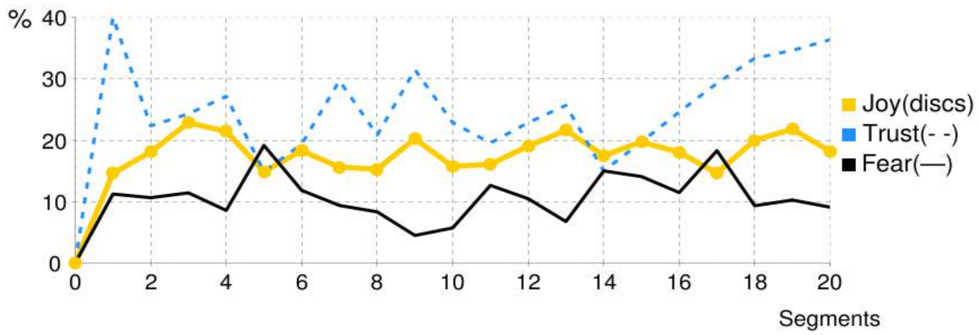
Tracking Emotions in Stories

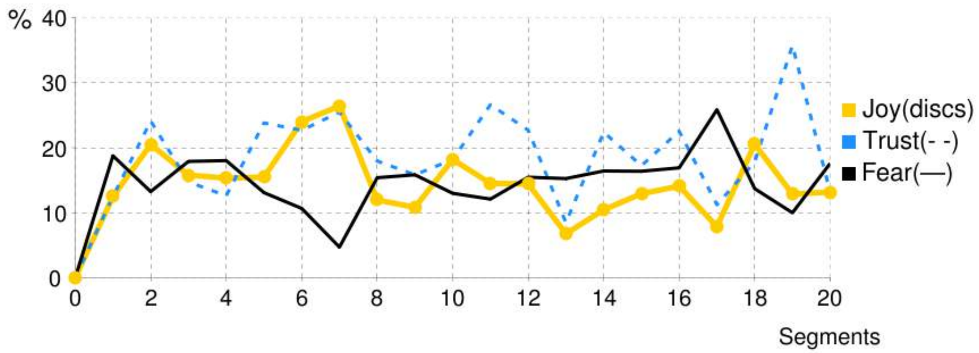
- Can we automatically track the emotions of characters?
- Are there some canonical shapes common to most stories?
- Can we track the change in distribution of emotion words?

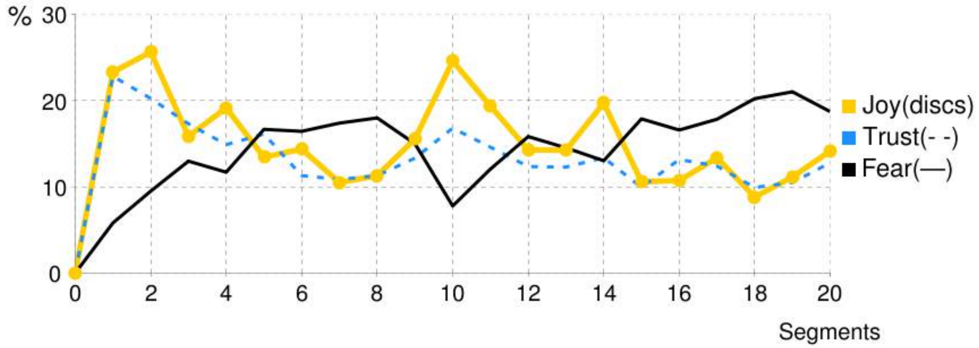


SOURCE DAVID YANG, VISUAL.LY

HBR.ORG


As You Like It


Hamlet


Frankenstein

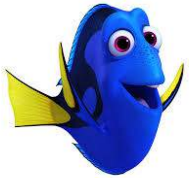
Work on shapes of stories

- 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.

- Character-based kernels for novelistic plot structure. Elsner, M., 2012, April. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 634-644). Association for Computational Linguistics.

- A novel method for detecting plot. M. Jockers http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/, June 2014.

- The emotional arcs of stories are dominated by six basic shapes. Reagan, A.J., Mitchell, L., Kiley, D., Danforth, C.M. and Dodds, P.S., 2016. EPJ Data Science, 5(1), p.31.

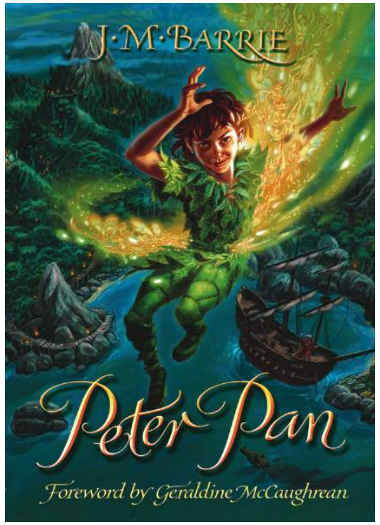


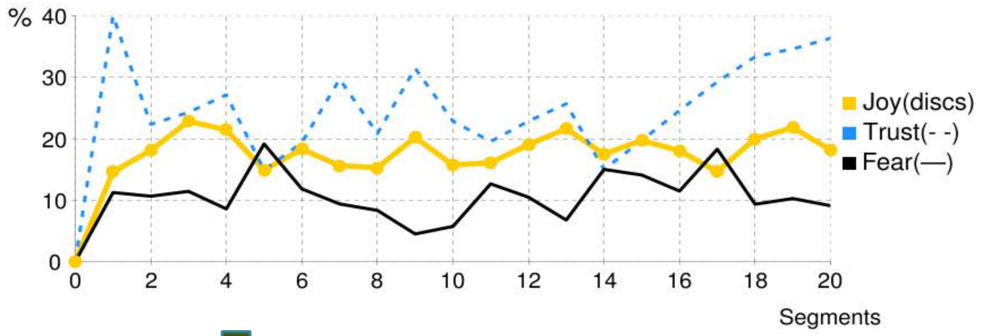
creativity

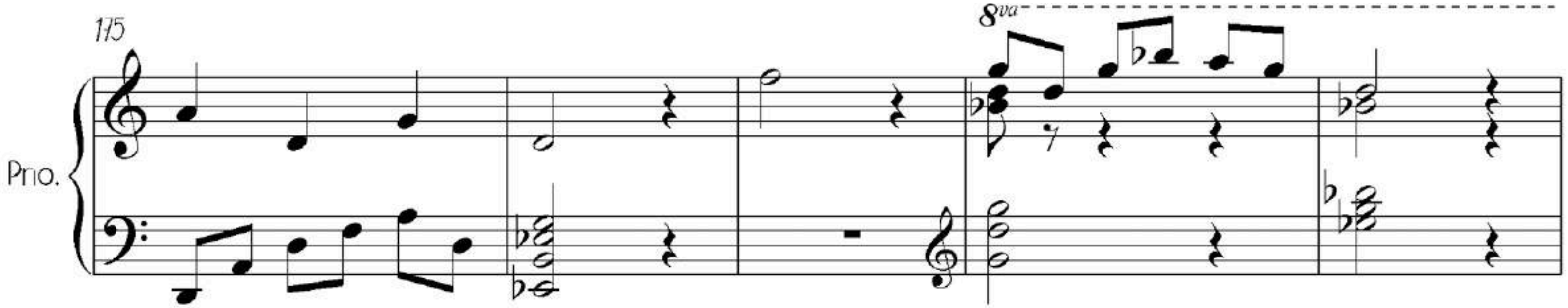
Generating music from text

Paper:

- Generating Music from Literature. Hannah Davis and Saif M. Mohammad, In Proceedings of the EACL Workshop on Computational Linguistics for Literature, April 2014, Gothenburg, Sweden.







A method to generate music from literature.

- music that captures the change in the distribution of emotion words.

Music-Emotion Associations

Hannah Davis
Artist/Programmer

- 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

Hunter et al., 2010, Hunter et al., 2008, Ali and Peynirciolu, 2010,
Gabrielsson and Lindstrom, 2001, Webster and Weir, 2005

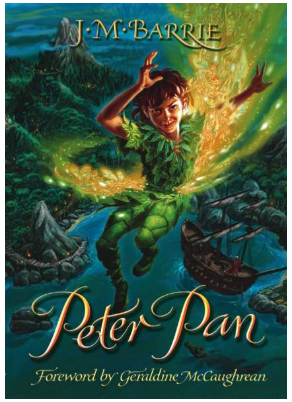
TransProse

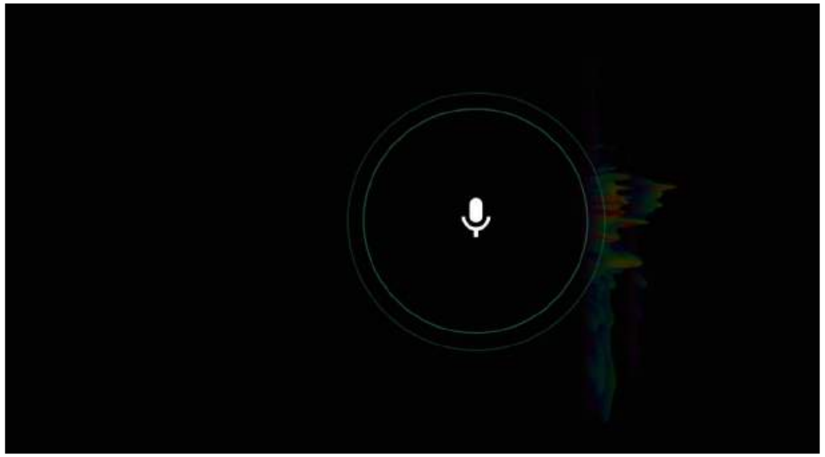
Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.

TransProse

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

Examples

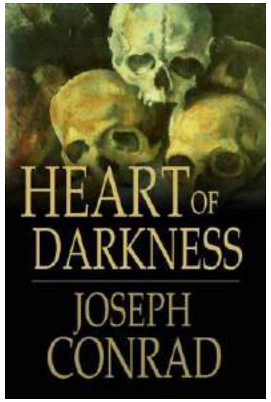




TransProse

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

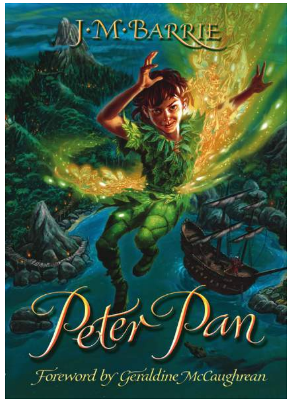
Examples

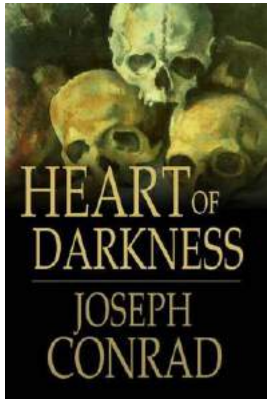


TransProse

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

Examples





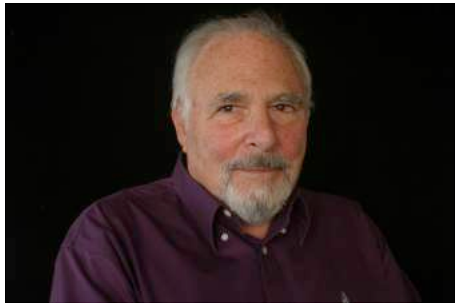
TransProse: www.musicfromtext.com
Music played 300,000 times since website launched in April 2014.

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

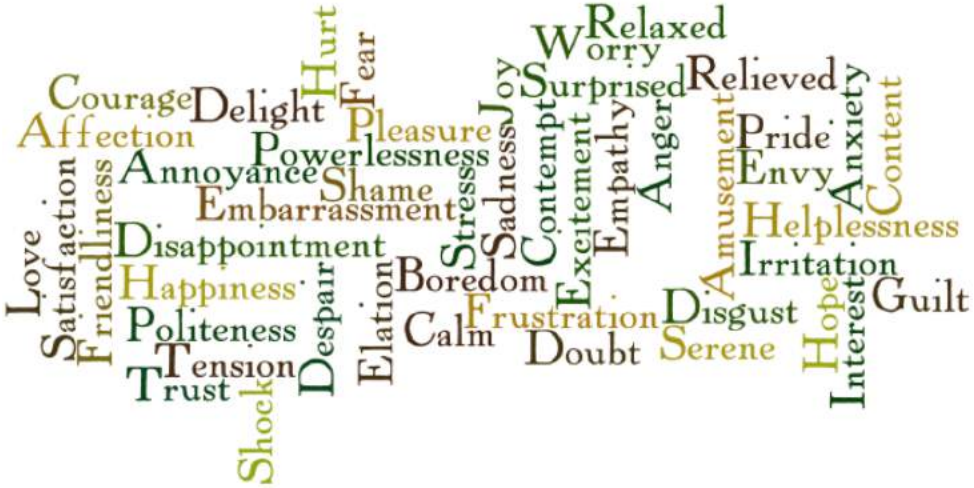
Margaret Mead
Cultural anthropologist



Paul Ekman
Psychologist and discoverer of micro expressions.

Lisa Barrett
University Distinguished Professor of Psychology, Northeastern University

- 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



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**

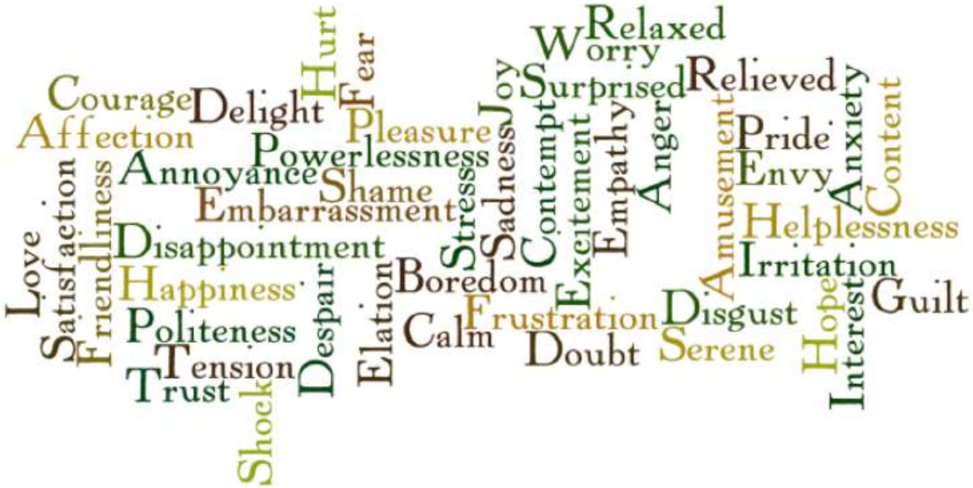
- Hashtags are not always good labels:
  - hashtag used sarcastically

    The reviewers want me to re-annotate the data. **#joy**

Paper:

#Emotional Tweets, Saif Mohammad, In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem), June 2012, Montreal, Canada.

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.
- Using Hashtags to Capture Fine Emotion Categories from Tweets. Saif M. Mohammad, Svetlana Kiritchenko, Computational Intelligence, Volume 31, Issue 2, Pages 301-326, May 2015.

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

Svetlana Kiritchenko
NRC

Xiaodan Zhu
NRC

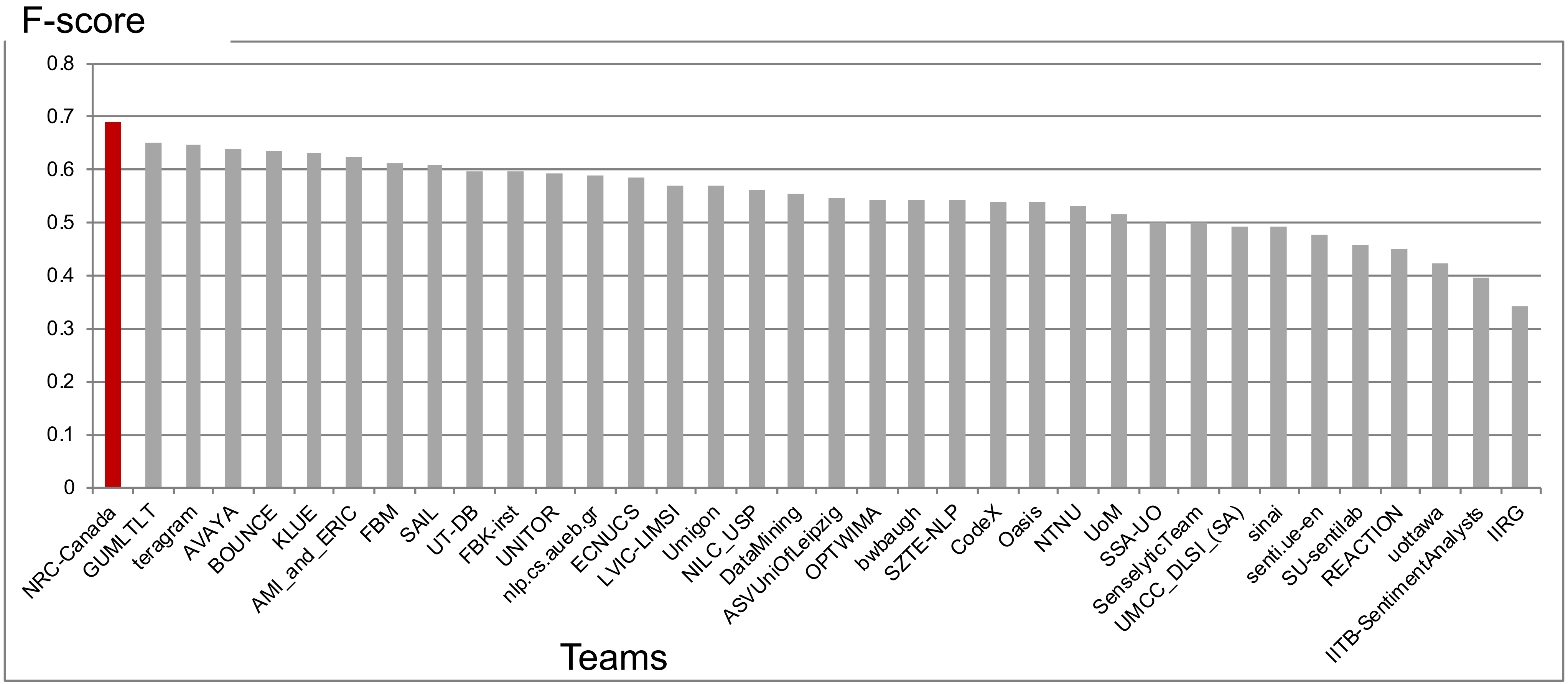
SemEval Shared task on the Sentiment Analysis of Tweets

Papers:

- Sentiment Analysis of Short Informal Texts. Svetlana Kiritchenko, Xiaodan Zhu and Saif Mohammad. *Journal of Artificial Intelligence Research*, 50, August 2014.
- 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



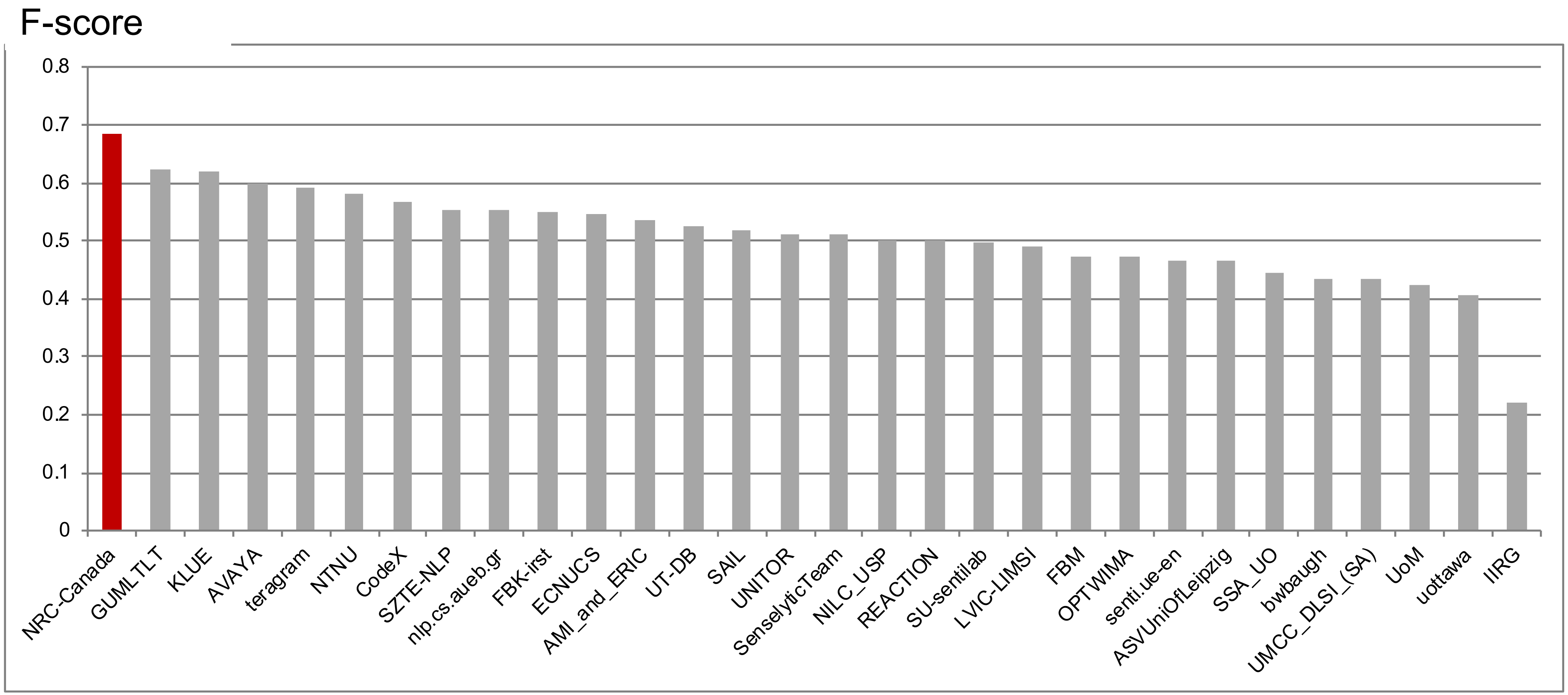
National Research Council Canada Conseil national de recherches Canada

@SaifMMohammad

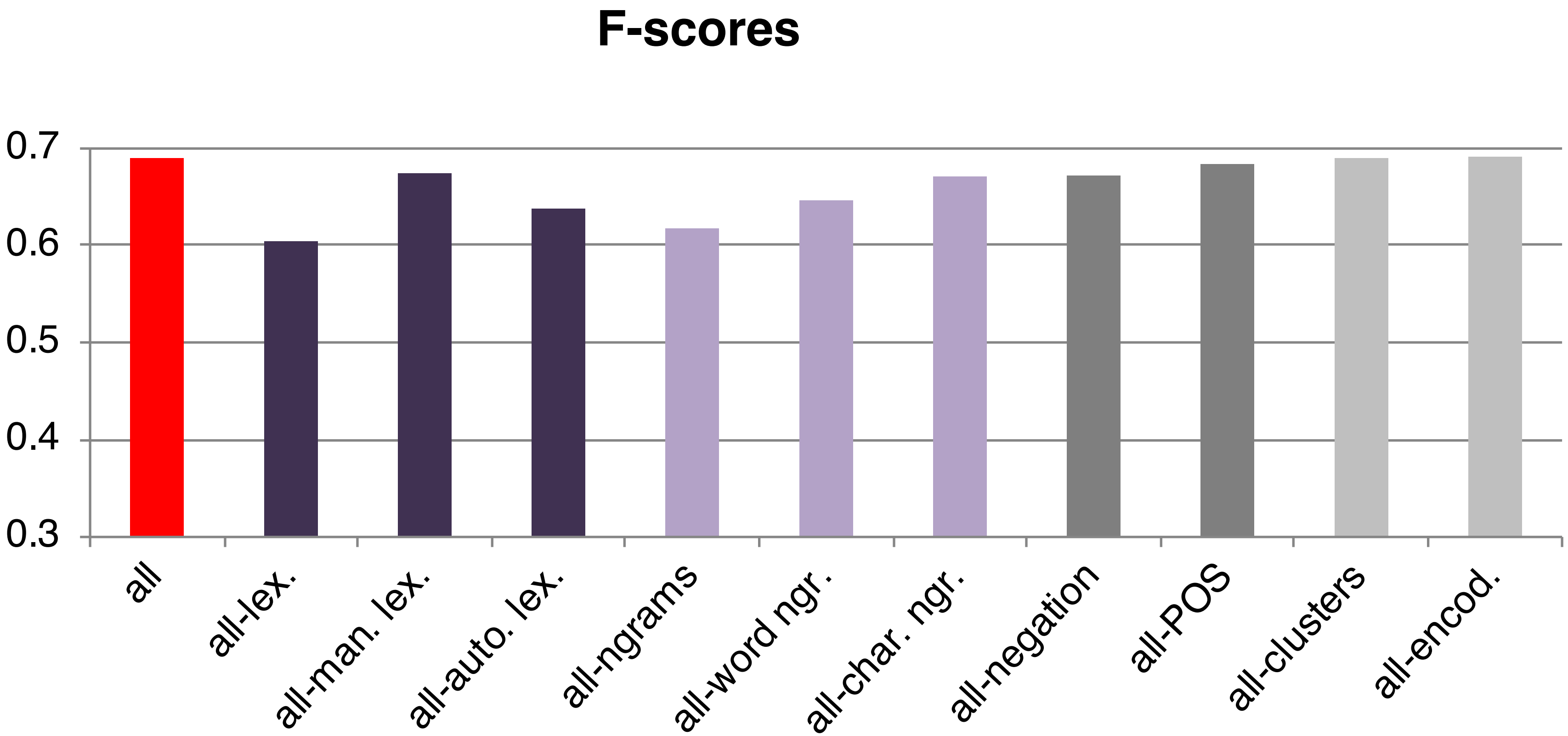
Canada

Sentiment Analysis Competition

SemEval-2013: Classify SMS messages, 30 teams



Feature Contributions (on Tweets)



Detecting Stance in Tweets

favor against neither

Parinaz Sobhani

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

Svetlana Kiritchenko

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.

Xiaodan Zhu

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.

Colin Cherry

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)

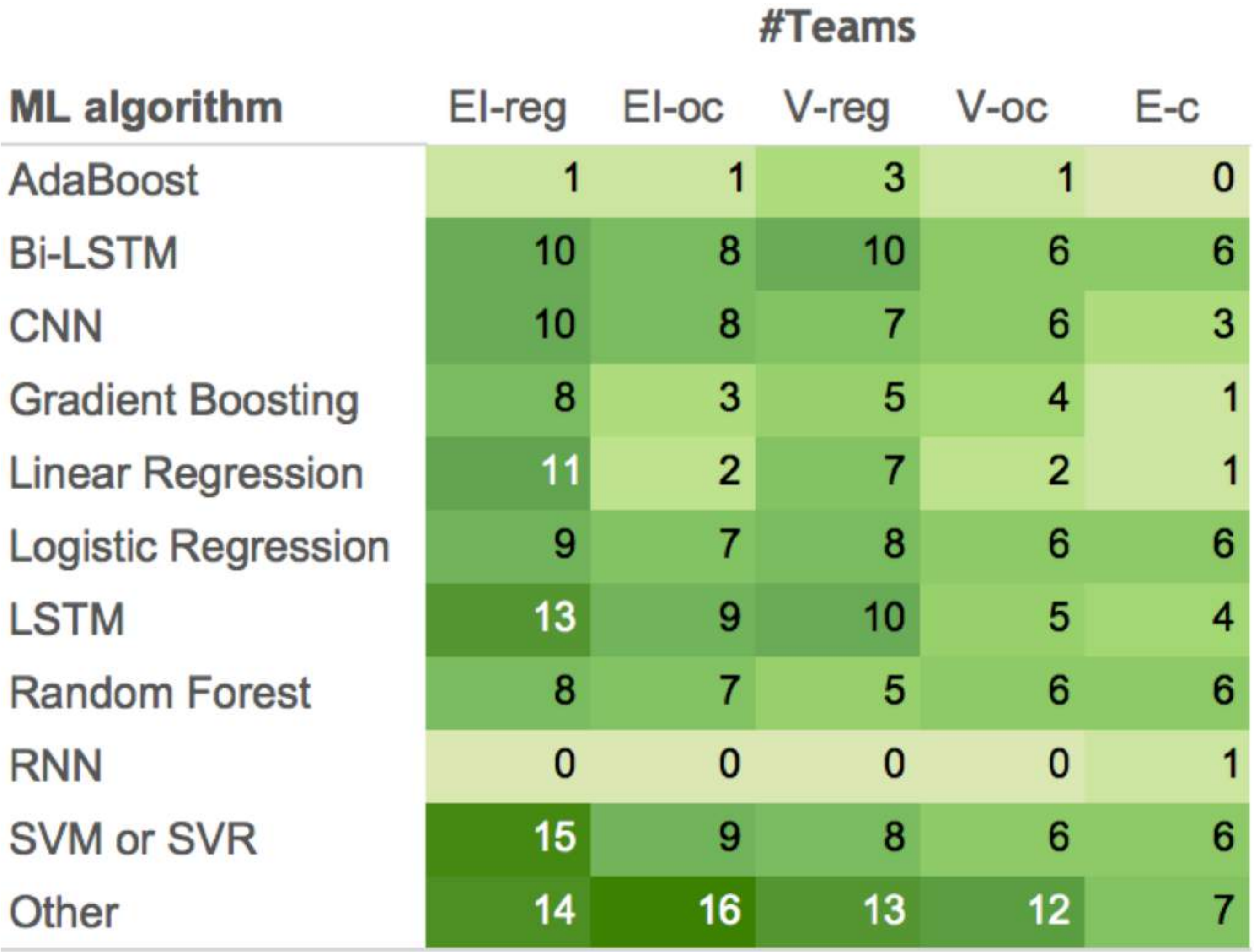
Felipe José Bravo Márquez

Mohammad Salameh

Svetlana Kiritchenko

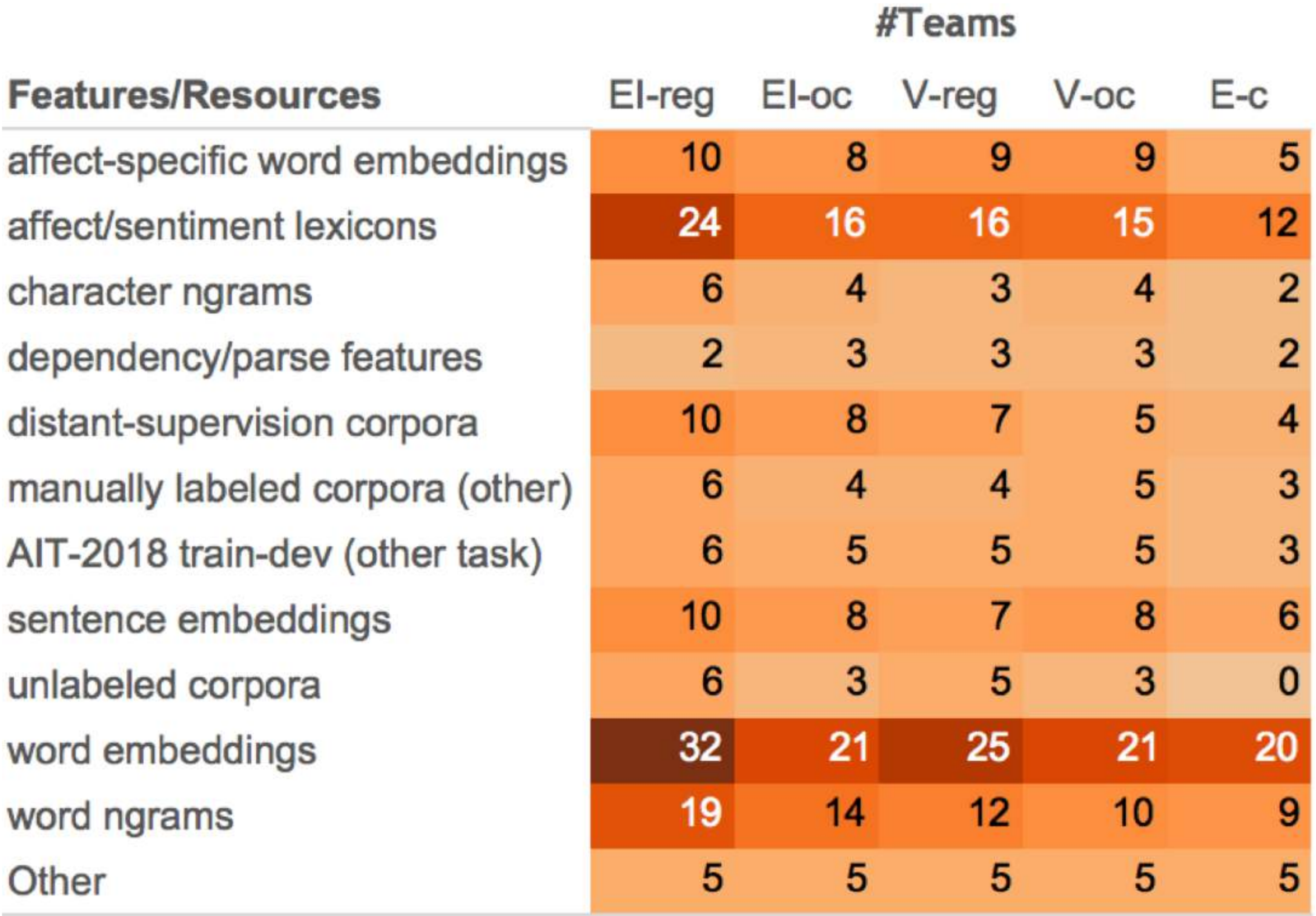
Semeval-2018 Task 1: Affect in tweets. Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. In Proceedings of International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, June 2018.

Participating Systems: ML algorithms



	#Teams				
ML algorithm	EI-reg	EI-oc	V-reg	V-oc	E-c
AdaBoost	1	1	3	1	0
Bi-LSTM	10	8	10	6	6
CNN	10	8	7	6	3
Gradient Boosting	8	3	5	4	1
Linear Regression	11	2	7	2	1
Logistic Regression	9	7	8	6	6
LSTM	13	9	10	5	4
Random Forest	8	7	5	6	6
RNN	0	0	0	0	1
SVM or SVR	15	9	8	6	6
Other	14	16	13	12	7

Participating Systems: features



	#Teams				
Features/Resources	EI-reg	EI-oc	V-reg	V-oc	E-c
affect-specific word embeddings	10	8	9	9	5
affect/sentiment lexicons	24	16	16	15	12
character ngrams	6	4	3	4	2
dependency/parse features	2	3	3	3	2
distant-supervision corpora	10	8	7	5	4
manually labeled corpora (other)	6	4	4	5	3
AIT-2018 train-dev (other task)	6	5	5	5	3
sentence embeddings	10	8	7	8	6
unlabeled corpora	6	3	5	3	0
word embeddings	32	21	25	21	20
word ngrams	19	14	12	10	9
Other	5	5	5	5	5

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)

fairness

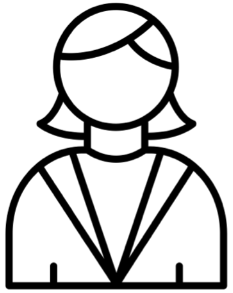
Includes a separate evaluation component for biases towards race and gender.

Do Machines Make Fair Decisions?

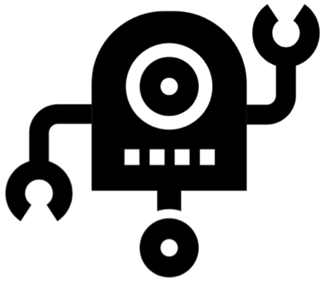
YES:

- 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



Created by Made
from Noun Project



Created by Oksana Latysheva
from Noun Project

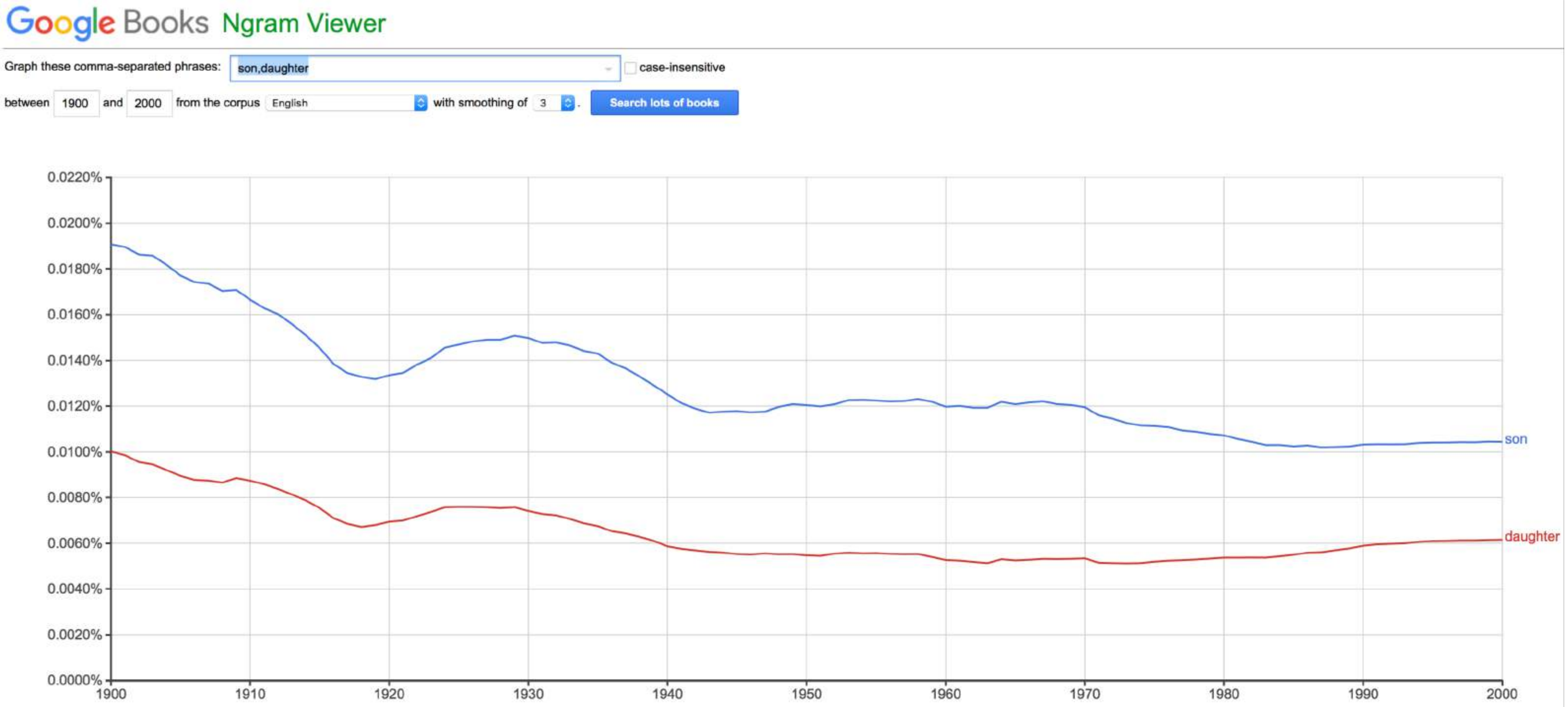
Do Machines Make Fair Decisions?

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

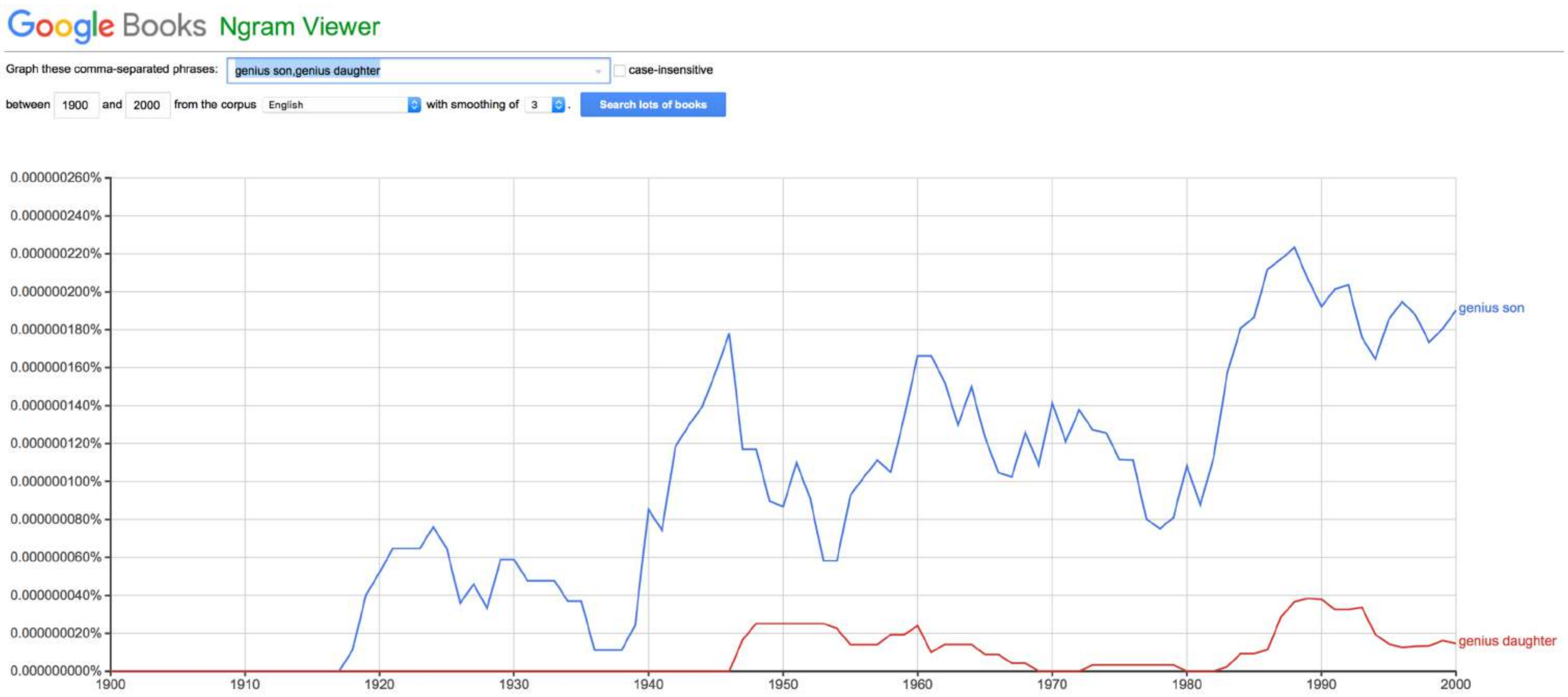
Occurrences of “son” and “daughter” in the Google Books Ngram corpus



Graph these comma-separated phrases: son,daughter □ case-insensitive

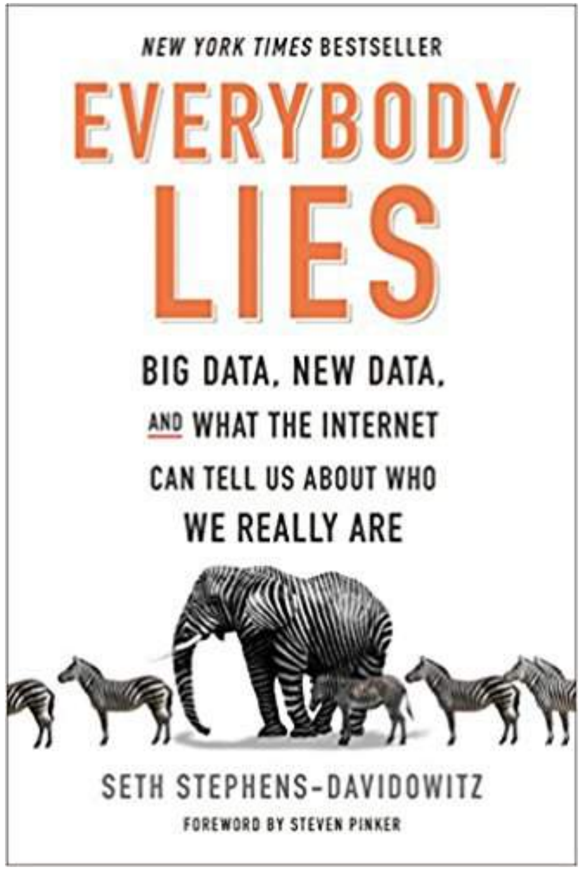
between 1900 and 2000 from the corpus English with smoothing of 3.

Occurrences of “genius son” and “genius daughter” in the Google Books Ngram corpus



Graph these comma-separated phrases: genius son,genius daughter □ case-insensitive

between 1900 and 2000 from the corpus English with smoothing of 3.



NEW YORK TIMES BESTSELLER

EVERYBODY LIES

BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE

SETH STEPHENS-DAVIDOWITZ

FOREWORD BY STEVEN PINKER

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

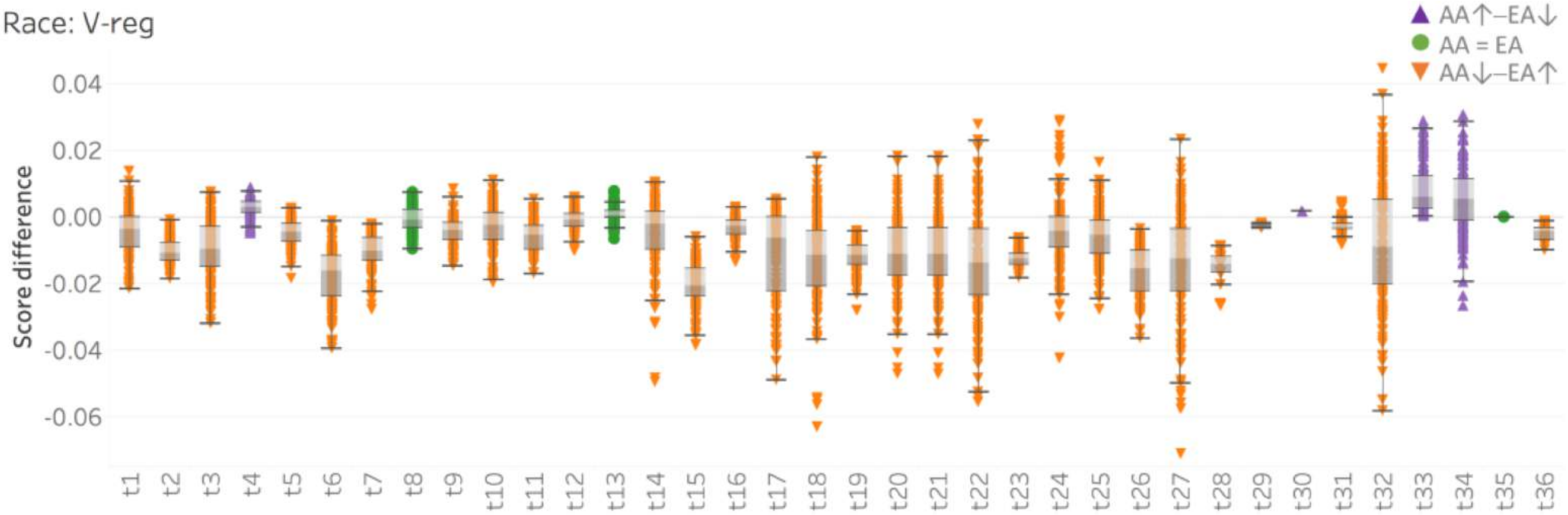
- focus on one or two systems or resources
    - word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Speer, 2017)
- no benchmark dataset for examining inappropriate biases

Svetlana Kiritchenko

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)



Race bias results on the Equity Evaluation Corpus.

Teams are ordered left to right by their performance on the Tweets Test Set (from best to worst).

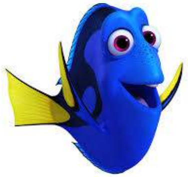
Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Svetlana Kiritchenko and Saif M. Mohammad. In *Proceedings of *Sem*, New Orleans, LA, USA, June 2018.

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

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Svetlana Kiritchenko and Saif M. Mohammad. In *Proceedings of *Sem*, New Orleans, LA, USA, June 2018.

Art and Emotions

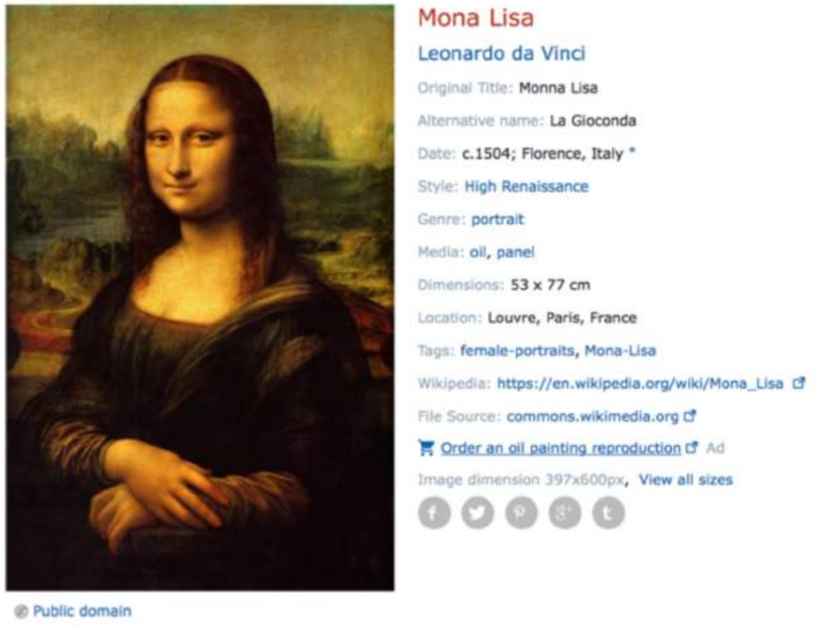


creativity

WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. Saif M. Mohammad and Svetlana Kiritchenko. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, May 2018, Miyazaki, Japan.

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



Mona Lisa

Leonardo da Vinci

Original Title: Monna Lisa

Alternative name: La Gioconda

Date: c.1504; Florence, Italy *

Style: High Renaissance

Genre: portrait

Media: oil, panel

Dimensions: 53 x 77 cm

Location: Louvre, Paris, France

Tags: female-portraits, Mona-Lisa

Wikipedia: https://en.wikipedia.org/wiki/Mona_Lisa

File Source: commons.wikimedia.org

Order an oil painting reproduction Ad

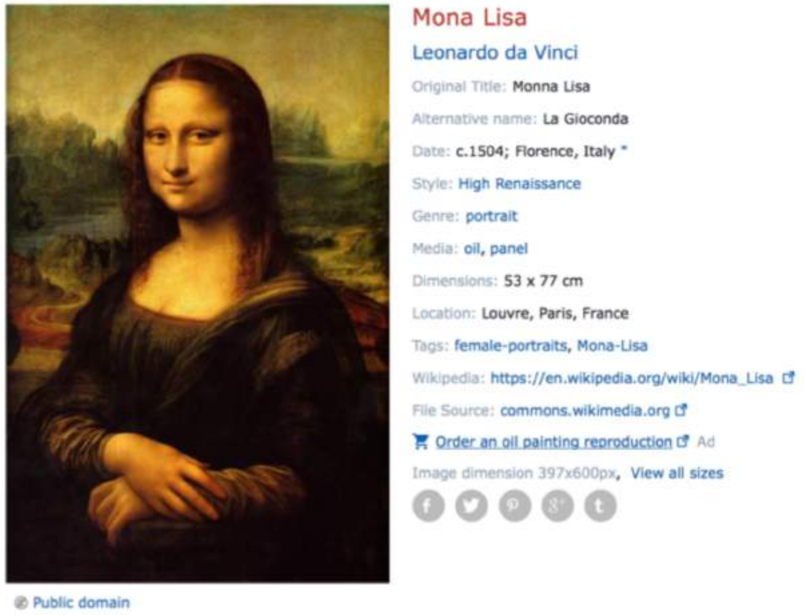
Image dimension 397x600px, View all sizes

Public domain

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).

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.



Mona Lisa

Leonardo da Vinci

Original Title: Monna Lisa

Alternative name: La Gioconda

Date: c.1504; Florence, Italy *

Style: High Renaissance

Genre: portrait

Media: oil, panel

Dimensions: 53 x 77 cm

Location: Louvre, Paris, France

Tags: female-portraits, Mona-Lisa

Wikipedia: https://en.wikipedia.org/wiki/Mona_Lisa

File Source: commons.wikimedia.org

Order an oil painting reproduction Ad

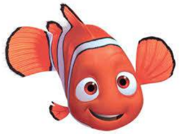
Image dimension 397x600px, View all sizes

Public domain

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 (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)

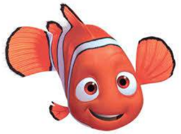
http://saifmohammad.com/WebPages/AffectIntensity.htm

The NRC Word–Colour Association Lexicon

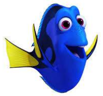
provides associations for ~14,000 words with 11 common colours

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

Summary



- Created several lexicons that capture word-emotion associations
- Used comparative annotations (best-worst scaling) to obtain reliable real-valued scores
- Showed new tasks and applications not just in sentiment analysis, but also in investigating how we use language
    - especially with regard to fairness and 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