SemEval-2018 Task 1: Affect in Tweets

Created by ATOM
from Noun Project

Saif M. Mohammad[1], Felipe José Bravo Márquez[2], Mohammad Salameh[3], Svetlana Kiritchenko[1]

[1]National Research Council Canada, [2]The University of Waikato, [3]Carnegie Mellon University, Qatar

Tasks (for English, Arabic, and Spanish Tweets)

1. Emotion Intensity Regression (EI-reg):

First introduced in the WASSA-2017 Shared Task: **Emotion Intensity in Tweets**

Given a tweet and an emotion E,

determine the intensity of E that best represents the mental state of the tweeter

- a real-valued score between 0 (least E) and 1 (most E)

Natural language applications benefit from knowing both the class of emotion and its intensity

- E.g., useful for commercial customer satisfaction system to distinguish between significant frustration or anger vs. instances of minor inconvenience

Tasks (for English, Arabic, and Spanish Tweets)

1. Emotion Intensity Regression (EI-reg):

   Given a tweet and an emotion E,

   determine the intensity of E that best represents the mental state of the tweeter

   - a real-valued score between 0 (least E) and 1 (most E)

2. Emotion Intensity Ordinal Classification (EI-oc):

   Given a tweet and an emotion E,

   classify the tweet into one of four ordinal classes of intensity of E that best represents the mental state of the tweeter

   - not angry, slightly angry, moderately angry, very angry

Tasks (for English, Arabic, and Spanish Tweets)

3. Valence (Sentiment) Regression (V-reg):

Given a tweet,

determine the intensity of sentiment or valence (V) that best represents the mental state of the tweeter

- a real-valued score between 0 (most negative) and 1 (most positive)

4. Valence Ordinal Classification (V-oc):

Given a tweet,

classify it into one of seven ordinal classes of valence (sentiment intensity) that best represents the mental state of the tweeter

- very negative, moderately negative, slightly negative, neutral or mixed, slightly positive, moderately positive, very positive

Tasks (for English, Arabic, and Spanish Tweets)

5. Emotion Classification (E-c):

Given a tweet,
classify it into one, or more, of twelve given categories
that best represent the mental state of the tweeter

- anger (also includes annoyance, rage)
- anticipation (also includes interest, vigilance)
- disgust (also includes disinterest, dislike, loathing)
- fear (also includes apprehension, anxiety, terror)
- joy (also includes serenity, ecstasy)
- love (also includes affection)
- optimism (also includes hopefulness, confidence)
- pessimism (also includes cynicism, no confidence)
- sadness (also includes pensiveness, grief)
- surprise (also includes distraction, amazement)
- trust (also includes acceptance, liking, admiration)
- neutral or no emotion

- Plutchik emotions
- other

Motivation

Human annotations of tweets for emotions 

- For use by automatic systems:
  - that detect emotions in tweets
  - other emotion related tasks such as detecting stance, personality traits, well-being, cyber-bullying, etc.

- To draw inferences about people:
  - to understand emotions, or how we convey emotions through language
  - how reliably we can order tweets as per emotion intensity
  - how the intensities of the basic emotions relate to valence

Collect Tweets using Query Terms

Query Terms:

- 50 to 100 related terms from the *Roget's Thesaurus*
  - associated with that emotion at different intensity levels
    - for anger: *angry, mad, frustrated, annoyed, peeved, irritated, miffed, fury,* and so on
    - for sadness: *sad, devastated, sullen, down, crying, dejected, heartbroken, grief,* and so on
- emojis and emoticons

Presence of terms does not guarantee an emotion or a certain intensity of the emotion.

- Overall, the set is relatively more likely to be conveying emotions

How to capture fine-grained emotion intensity reliably? A harder task!

Humans are not good at giving real-valued scores:

- 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

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

Intensity Annotations

Best–Worst Scaling (Louviere & Woodworth, 1990)**:**
Give k terms and ask which is most X, and which is least X
(*k is usually 4 or 5*)

- 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)

Example BWS Annotation Instance: for emotion intensity from tweets

Speaker 1: These days I see no light. Nothing is working out #depressed
Speaker 2: The refugees are the ones running from terror.
Speaker 3: Tim is sad that the business is not going to meet expectations.
Speaker 4: Too many people cannot make ends meet with their wages.

Q1. Which of the four speakers is likely to be the MOST SAD (or having a mental state most inclined towards sadness)

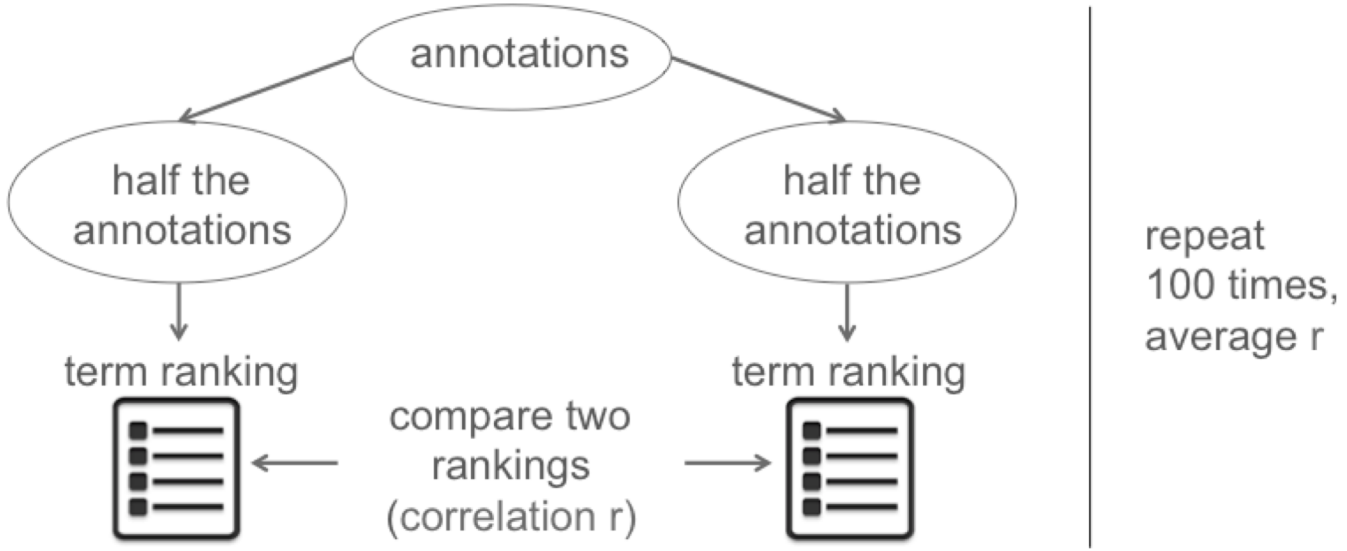
Q2. Which of the four speakers is likely to be the LEAST SAD (or having a mental state least inclined towards sadness)

Once annotations are done:

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

Reliability (Reproducibility) of Annotations

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



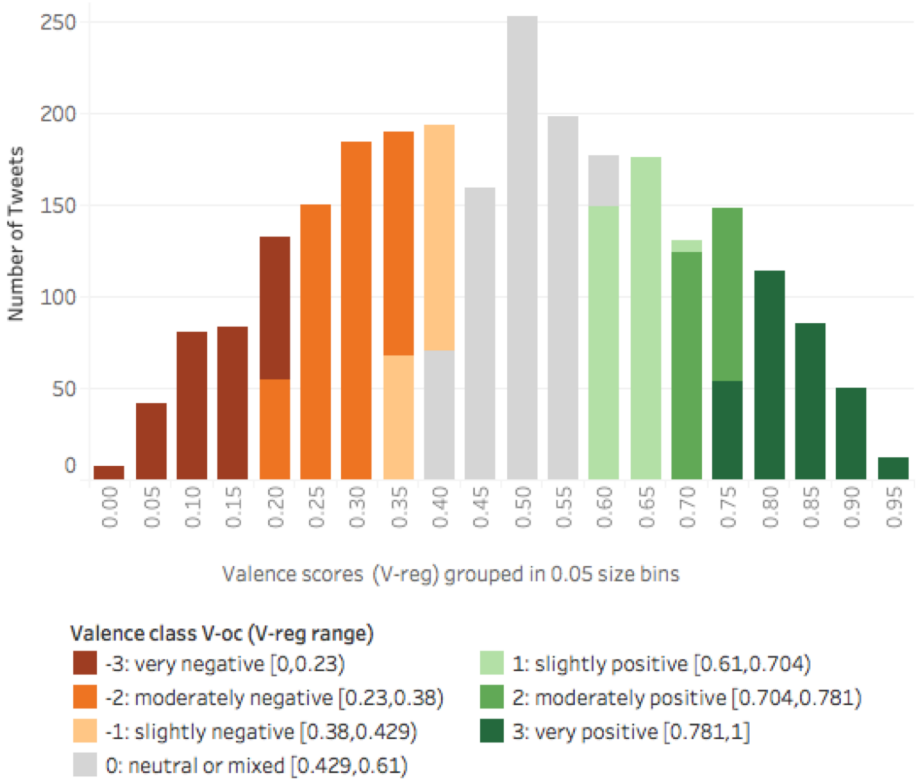
Split-Half Reliability: Emotion Intensity Annotations

Emotion	Spearman Corr. (r)	Pearson Corr. (ρ)
anger	0.89	0.90
fear	0.84	0.85
joy	0.90	0.91
sadness	0.82	0.83
valence	0.92	0.92

High correlation numbers indicate a high degree of reproducibility.

Similar split-half reliabilities for Arabic and Spanish annotations.

Distribution: Valence score (V-reg) and Valence class (V-oc)



The boundaries between valence classes were manually identified by the task organizers.

Affect in Tweets Dataset

Annotated Data:

- 1,400 to 11,000 tweets per task-language pair
- split into train, dev, and test sets

Official evaluation metrics:

- EI-reg, EI-oc, V-reg, and V-oc:
  - Pearson Correlation Coefficient
- E-c:
  - multi-label accuracy or Jaccard index (size of the intersection of the predicted and gold label sets divided by the size of their union)

Dataset	Total
English	
E-c	10,983
EI-reg, EI-oc	
anger	3,091
fear	3,627
joy	3,011
sadness	2,905
V-reg, V-oc	2,567
Arabic	
E-c	4,381
EI-reg, EI-oc	
anger	1,400
fear	1,400
joy	1,400
sadness	1,400
V-reg, V-oc	1,800
Spanish	
E-c	7,094
EI-reg, EI-oc	
anger	1,986
fear	1,986
joy	1,990
sadness	1,991
V-reg, V-oc	2,443

72 Teams, 319 systems

#systems in each task--language pair:

Task	English	Arabic	Spanish	All
EI-reg	48	13	15	76
EI-oc	37	12	14	63
V-reg	37	13	13	63
V-oc	35	13	12	60
E-c	33	12	12	57
Total	190	63	66	319

Result Tables: 5 tasks * 3 languages = 15

Results Summary

- Best results for English Tasks:
  - SeerNet (EI-reg: ~80, EI-oc: ~70, V-reg: ~87, V-oc: ~84)
  - NTUA-SLP (E-c: ~59)

  Often ~30 points higher than the unigrams baseline
- Best results for Arabic Tasks:
  - AffectThor (EI-reg, EI-oc), EiTAKA (V-reg, V-oc), EMA (E-c)
- Best results for Spanish Tasks:
  - AffectThor (EI-reg, EI-oc, V-reg), Amobee (V-oc), MILAB-SNU (E-c)

Results for Arabic and Spanish are lower than that for English.

Further details on the 15 result tables are in the paper.

Official evaluation metric for EI-reg, EI-oc, V-reg, and V-oc was Pearson Correlation Coefficient; and for E-c was multi-label accuracy (or Jaccard index).

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

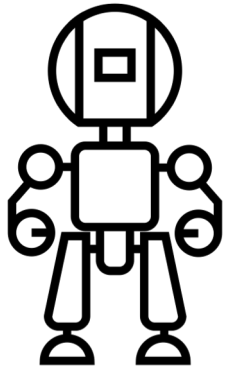
	#Teams				
Features/Resources	EI-reg	EI-oc	V-reg	V-oc	E-c
affect-specific word embeddings	10	8	9	9	5
affect/sentiment lexicons	24	16	16	15	12
character ngrams	6	4	3	4	2
dependency/parse features	2	3	3	3	2
distant-supervision corpora	10	8	7	5	4
manually labeled corpora (other)	6	4	4	5	3
AIT-2018 train-dev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

Participating Systems: Affect Lexicons

Lexicon	#Teams
AFINN	23
ANEW	9
Arabic translation of the NRC Emotion Lexicon	4
Bing Liu Lexicon	23
ElhPolar polarity lexicon for Spanish	3
LIWC	5
Mohammad et al.'s Arabic Emoticon Lexicon	5
Mohammad et al.'s Arabic Hashtag Lexicon	5
Mohammad et al.'s Arabic Hashtag Lexicon (dialectal)	2
MPQA	21
NRC Affect Intensity Lexicon	21
NRC Emoticon Lexicon (Sentiment140)	24
NRC Emotion Lexicon (EmoLex)	22
NRC Hashtag Emotion Lexicon	23
NRC Hashtag Sentiment Lexicon	25
SentiStrength	18
SentiWordNet	18
Spanish translation of the NRC Emotion Lexicon	5
No lexicons used	29

Examining Bias

in Sentiment Analysis Systems



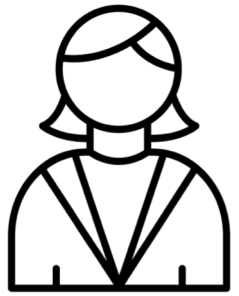
Created by iconcheese
from Noun Project

Do Machines Make Fair Decisions?

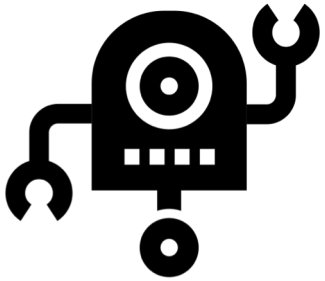
YES:

- 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



Created by Made from Noun Project



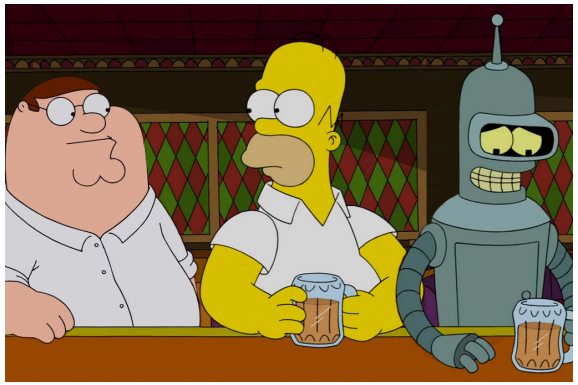
Created by Oksana Latysheva from Noun Project

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



Our Goal:

- Measure the extent to which systems consistently assign higher/lower scores to sentences mentioning one gender/race compared to another gender/race

Our Approach:

- Equity Evaluation Corpus (EEC)—a dataset of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders
- Examine the output of 219 sentiment analysis systems
  - compare emotion and sentiment intensity scores on pairs of sentences that differ only in one word corresponding to race or gender

    This man made me feel angry vs. This woman made me feel angry

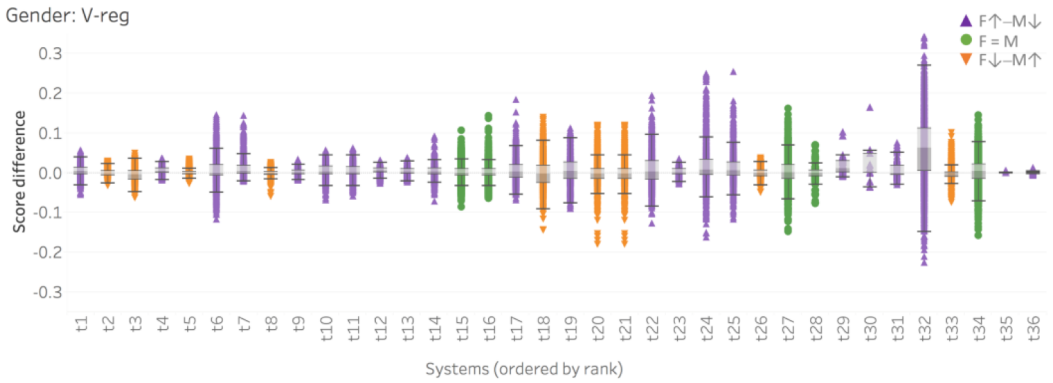
Results: Bias in Systems

- Common
  - ~75% of the systems consistently mark sentences involving one gender/race with higher intensity scores
- More common for race than for gender
- Different depending on the affect dimension involved

Created by Sean Maldjian from Noun Project

*Sem Talk 2pm today!

(Strand 12B)



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

Summary

Introduced an array of tasks where automatic systems have to infer the affectual state of a person from their tweet:

- 11 tasks, three languages
- new Affect in Tweets Dataset
  - more than 22,000 tweets annotated for coarse classes and for fine-grained affect labels/scores
- an evaluation component for measuring biases in the systems
- 72 teams (~200 participants), using a variety of ML architectures and resources

Beyond the shared task:

- developing better machine learning algorithms for detecting affect
- understanding emotions and relations between affect categories

Created by Symbolon from Noun Project

Resources Available at: www.saifmohammad.com

- Affect in Tweets Data
- Emotion and Sentiment Lexicons
- Links to Shared Tasks
- Interactive Visualizations

Contact:
saif.mohammad@nrc-cnrc.gc.ca

@SaifMMohammad

Saif M, Mohammad, Felipe José Bravo Márquez, Mohammad Salameh, Svetlana Kiritchenko