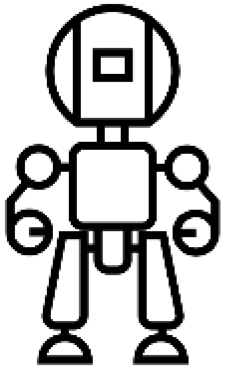


Created by iconcheese
from Noun Project

Examining Gender and Race Bias

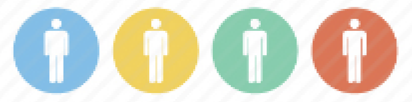
in Two Hundred Sentiment Analysis Systems

Svetlana Kiritchenko

Research Scientist, National Research Council Canada

(joint work with Saif M. Mohammad)

My Work: Natural Language Processing



Human annotations of words, phrases, tweets for sentiment and emotions

- Creating lexicons of words and phrases associated with emotions
- Draw inferences about language and people:
  - understand how we (or different groups of people) use language to express meaning and emotions

Develop automatic systems

- Predicting emotions of words, tweets, sentences, etc.
- Detecting stance, personality traits, adverse drug reactions, etc.

My Work: Natural Language Processing

Human annotations of words, phrases, tweets for sentiment and emotions

- Creating Sentiment Composition lexicons
  - real-valued sentiment association scores for phrases and their constituents
- Annotating tweets for emotion intensities and sentiment (valence)
  - analyzing the relations between different models of emotions (basic emotions vs. valence-arousal-dominance)
- Annotating tweets for stance and sentiment
  - stance: the tweeter is in favor of, against, or neutral towards a given target
  - target may not be explicitly mentioned in the text
  - analyzing the relationships between stance and sentiment

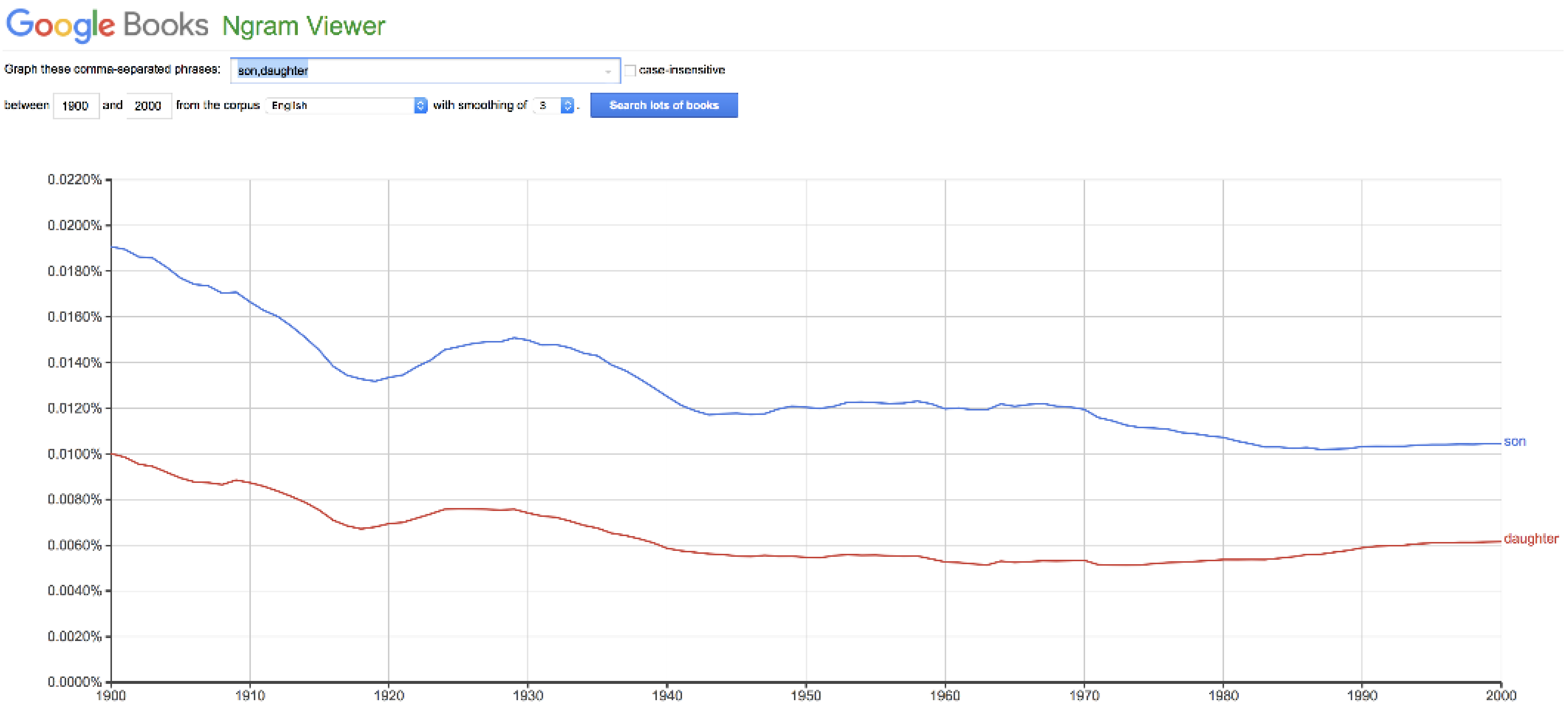
My Work: Natural Language Processing

Develop automatic systems

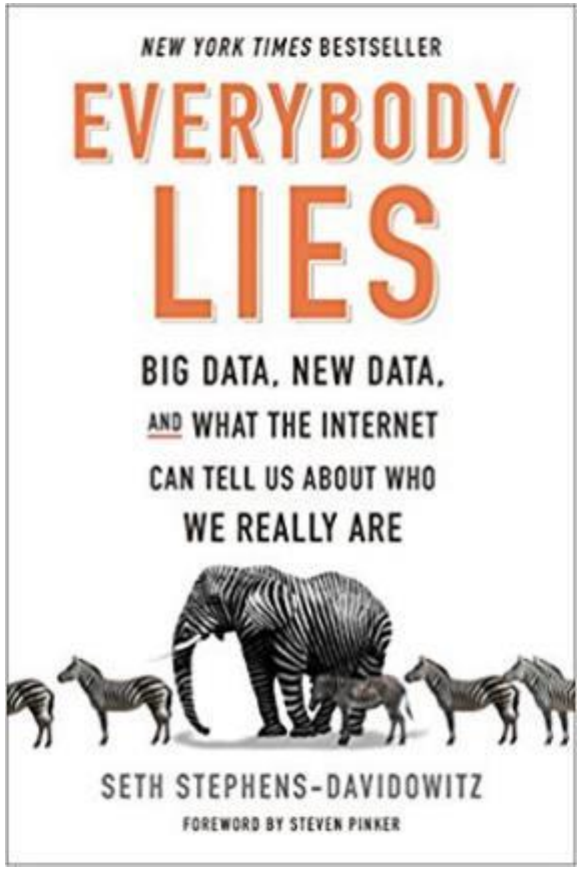
- Generating large-scale sentiment lexicons for words and phrases from tweets
  - using hashtags and emoticons
- Competition-winning text classification systems:
  - sentiment classification of tweets (SemEval-2013 and SemEval-2014)
  - aspect based sentiment analysis (SemEval-2014)
  - stance detection (SemEval-2016)
  - classification of tweets mentioning adverse drug reactions (AMIA-2017 Shared Task on Social Media Mining for Health Applications)

Biases in Language

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?

Inappropriate Human Biases

- Search the web disproportionally for intelligence and appearance related concepts for males and females, respectively
- Prefer CVs with white sounding names vs. African American or Hispanic names
- Give more frequent promotions and pay hikes to men vs. women
- Portray female and non-white characters with less agency and in negative light in books and movies

Machine Biases

Recent studies have demonstrated that predictive models built on historical data may inadvertently inherit inappropriate human biases.

Examples:

- loan eligibility and crime recidivism prediction systems that negatively assess people belonging to a certain zip code (Chouldechova, 2017)
- resumé sorting systems that believe that men are more qualified to be programmers than women (Bolukbasi et al., 2016)

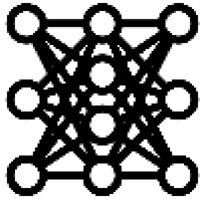
What We Can Do

Know the System Biases:

- what biases exist
  - e.g., whether predictions change depending on the gender or race of the person mentioned
- the extent of biases
- which biases are inappropriate

Where Necessary, Address the System Biases:

- explain system predictions
  - e.g., the extent to which a race or gender bias contributed to the decision



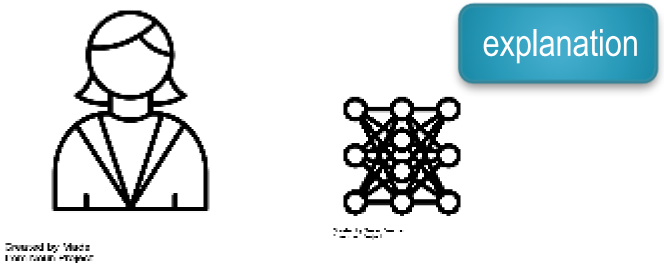
What We Can Do

Know the System Biases:

- what biases exist
  - e.g., whether predictions change depending on the gender or race of the person mentioned
- the extent of biases
- which biases are inappropriate

Where Necessary, Address the System Biases:

- explain system predictions
  - e.g., the extent to which a race or gender bias contributed to the decision
- algorithmically remove certain biases



Know the System Biases: Past Work

Focus on one or two systems or resources

- word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Speer, 2017)

Not many benchmark datasets for examining inappropriate biases in NLP systems

Know the System 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

Sentiment/Emotion Task

Given a tweet and an emotion E (say, anger):

- determine the intensity of emotion E that best represents the mental state of the tweeter
  - a real-valued score between 0 (least E) and 1 (most E)

That jerk stole my photo on tumblr #grrrr	anger score: 0.56
Some idiot put a major dent on my new car and did not even bother to leave his number! So pissed!! #$@!!2%&	anger score: 0.83

Goal of the Bias Examination:

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

Our Approach:

- compare emotion and sentiment intensity scores that the systems predict on pairs of sentences in the EEC that differ only in one word corresponding to race or gender

  This <u>man</u> made me feel angry vs. This <u>woman</u> made me feel angry

Here *system* refers to the combination of a machine learning architecture trained on a labeled dataset, and possibly using additional language resources

- bias can originate from any or several of these parts

This work examines the predictions of automatic systems—how they are.
We want to encourage deliberation and discussion on—how they should be and how to get there.

Disclaimer

- EEC is not a catch-all for all inappropriate biases
  - just one of several ways by which we can examine the fairness of sentiment analysis systems
- Any such mechanism is liable to be circumvented
- There are no simple solutions for comprehensively dealing with inappropriate human biases

Equity Evaluation Corpus

Problems with Taking Sentences from the Wild

- Hard to find pairs of sentences that differ in just one race or gender word
- Taking a sentence from the wild and switching the race or gender word is often not sufficient—context matters:
  - just switching the name or gender may not lead to a natural sentence
  - changing gender will involve tracking pronouns
  - multiple people may be mentioned (possibly with pronouns)
  - many of the names in social media and news articles refer to celebrities
- The sentences are often complex
  - may involve several emotions
  - expressed by and towards different entities

Nonetheless, an interesting avenue for future research.

The Equity Evaluation Corpus

Sentences with the following properties:

- include at least one gender- or race-associated word
- short and grammatically simple
- some sentences include explicit expressions of sentiment or emotion

Equity Evaluation Corpus

We created simple templates and generated sentences from them:

- Seven templates with emotion words, such as

  <Person> feels <emotion word>.
  She feels sad.

  <Person> found himself/herself in a/an <emotion word> situation.

  Latisha found herself in a terrifying situation.

- Four templates with no emotion words (neutral sentences), such as

  I talked to <person> yesterday.

  I talked to my mom yesterday.

The variables <person> and <emotion word> are replaced with pre-chosen values.

Equity Evaluation Corpus

<Person> is instantiated with one of the following values:

- ten common African American female first names

  Latisha feels sad.

- ten common African American male first names

  Jamel feels sad.

- ten common European American female first names

  Melanie feels sad.

- ten common European American male first names

  Harry feels sad.

- ten pairs of corresponding noun phrases referring to females and males

  My daughter feels sad.

  My son feels sad.

(names from Caliskan et al., 2017)

Equity Evaluation Corpus

<Emotion word> values:

- ten words for each of the four basic emotions: anger, fear, joy, sadness
- chosen from Roget's Thesaurus
- examples: *sad, scared, devastated, happy,* etc.

The eleven templates along with <person> and <emotion word> values led to 8,640 sentences in total.

The Sentiment Analysis Task

SemEval-2018 Affect in Tweets Shared Task

Tasks:

1. Emotion Intensity Regression (EI-reg):

    Given a tweet and an emotion E (anger, fear, joy, or sadness),

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

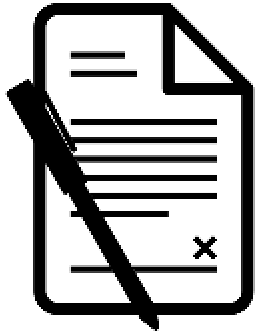
    Given a tweet, determine the intensity of sentiment or valence (V) that best represents the mental state of the tweeter
    – a real-valued score between 0 (most negative) and 1 (most positive)

For each task, there were two kinds of test sets:

- regular tweets test sets
- EEC sentences (mystery test set)

First Bullet in the Terms and Conditions of the Shared Task

- By submitting results to this competition, you consent to the public release of your scores at this website and at SemEval-2018 workshop and in the associated proceedings, at the task organizers' discretion. Scores may include, but are not limited to, automatic and manual quantitative judgements, qualitative judgements, and such other metrics as the task organizers see fit. You accept that the ultimate decision of metric choice and score value is that of the task organizers.



Created by Douglas Santos
from Noun Project

Participating Systems

Fifty teams participated in one or both of the tasks – 219 submissions

- machine learning algorithms:
  - deep neural networks (LSTM, Bi-LSTM, etc.)
  - traditional (SVM/SVR, Logistic Regression, etc.)

  ML algorithms may accentuate or minimize biases in the data.

- features:
  - word embeddings
  - word ngrams
  - deep neural representations of tweets (sentence embeddings) trained on:
    - the provided training data
    - other manually labeled sentiment corpora
    - distant supervision corpus (provided by the task and other corpora)
  - features derived from existing sentiment and emotion lexicons

  All of these features are from resources that are potential sources of bias.

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

Features/Resources	#Teams 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

Quantifying Bias

Measuring Gender Bias

Each system predicts emotion intensity scores for each EEC sentence.

For each template, we compare:

- scores for sentence pairs that differ in noun phrase

  Δ = score(The conversation with my mom was heartbreaking) –

  score(The conversation with my dad was heartbreaking)

- average score for the set of sentences with female names with the average score for the set of sentences with male names

  Δ = average score(The conversation with <female name> was heartbreaking) –

  average score(The conversation with <male name> was heartbreaking)

1,584 pairs of scores are compared

Measuring Gender Bias

Three groups of systems:

- F = M: no statistically significant difference in intensity scores predicted for corresponding female and male sentences
- F↑–M↓: consistently gave higher scores for female sentences than for corresponding male sentences
- F↓–M↑: consistently gave lower scores for female sentences than for corresponding male sentences

Statistical significance: paired t-test (significance level of 0.05) with Bonferroni correction

Gender Bias Results

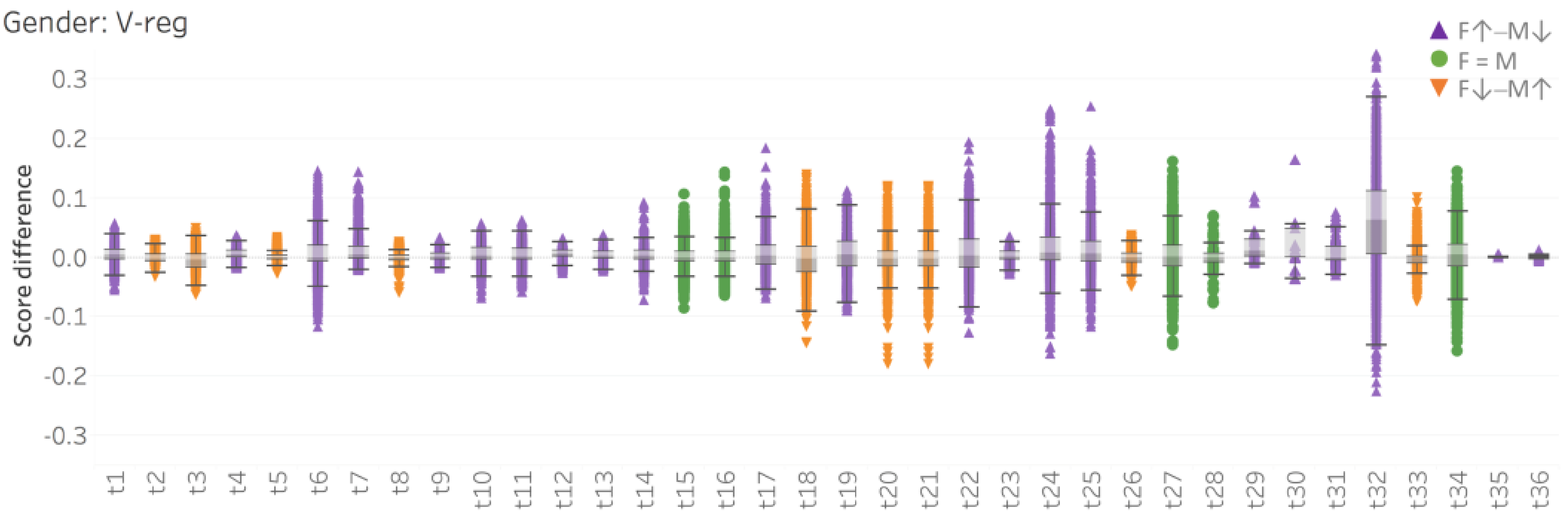
Created by Sean Maldjian
from Noun Project

The number of systems in each group:

Task	F = M	F↑–M↓	F↓–M↑	all
EI-reg				
anger	12	21	13	46
fear	11	12	23	46
joy	12	25	8	45
sadness	12	18	16	46
V-reg	5	22	9	36

- no statistically significant score difference:
  - only ~25% of the systems on the emotion tasks
  - only ~14% on the valence task
- systems tend to give higher scores to:
  - female sentences when predicting anger, joy, or valence
  - male sentences when predicting fear

Gender Bias Results: Box plot of the score differences on the gender sentence pairs for each system on the valence regression task (plots for the four emotions are similar)



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).

- Systems that showed no bias (shown in green) were also the teams that performed poorly on the tweets test set
- Average(|Δ|) < 0.03 (3%) on the [0,1] range
  - most of the score differences are small (50% of the points are in the grey box)

Measuring Race Bias

Three groups of submissions:

- AA = EA: no statistically significant difference in intensity scores predicted for sentences with African American (AA) and European American (EA) names
- AA↑–EA↓ : consistently gave higher scores for sentences with AA names than for corresponding sentences with EA names
- AA↓–EA↑: consistently gave lower scores for sentences with AA names than for corresponding sentences with EA names

Statistical significance: paired t-test (significance level of 0.05) with Bonferroni correction

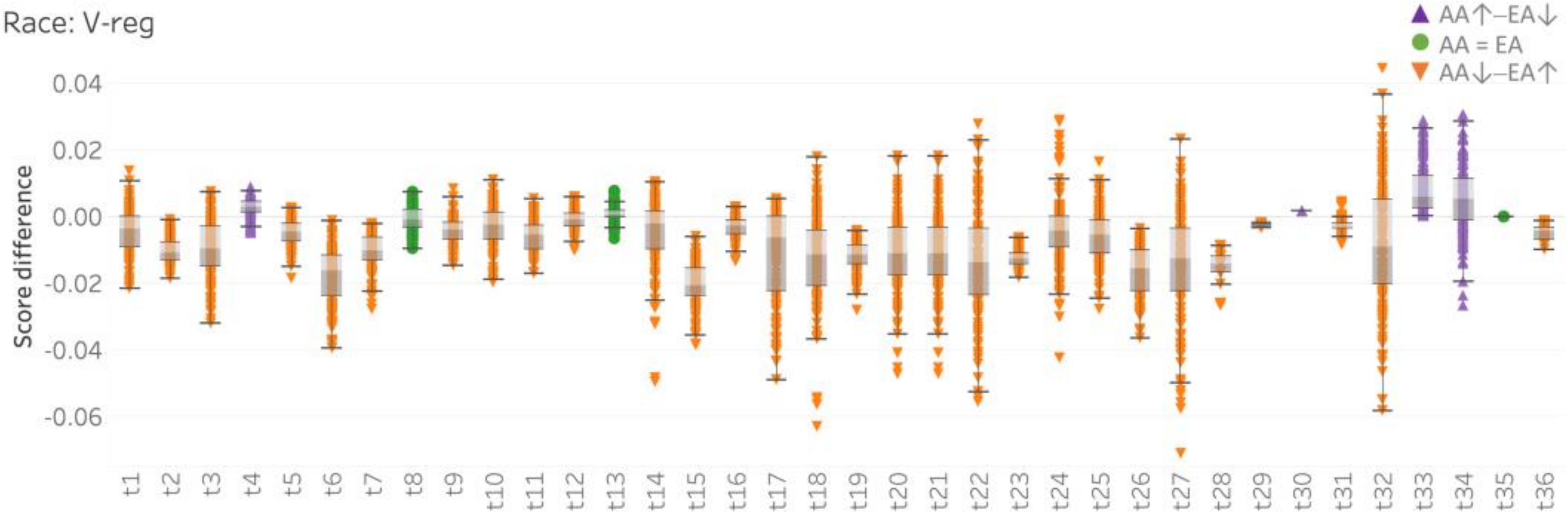
Race Bias Results

The number of systems in each group:

Task	AA = EA	AA↑–EA↓	AA↓–EA↑	All
EI-reg				
anger	11	28	7	46
fear	5	29	12	46
joy	8	7	30	45
sadness	6	35	5	46
V-reg	3	4	29	36

- no statistically significant score difference:
  - only 11-24% of the systems on the emotion tasks
  - only ~14% on the valence task
- systems tend to give higher scores to:
  - sentences with African American names when predicting anger, fear, or sadness
  - sentences with European American names when predicting joy or valence

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.

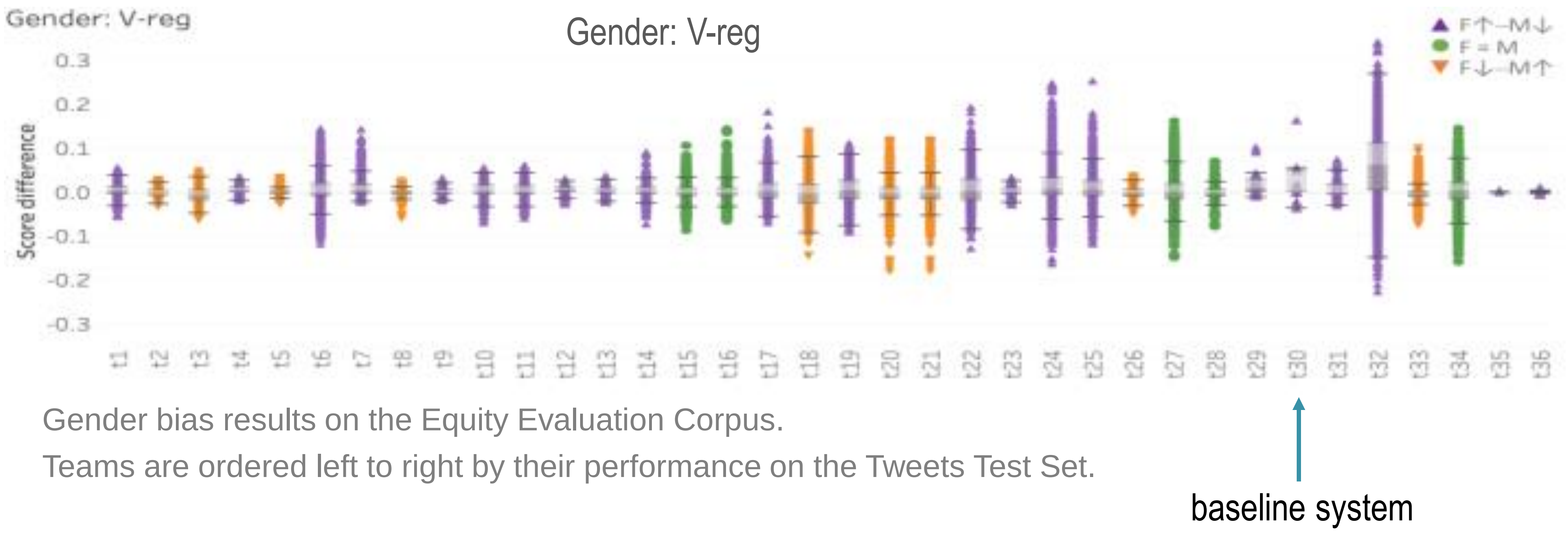
Teams are ordered left to right by their performance on the Tweets Test Set (from best to worst).

- Δ values spread over smaller intervals than on gender sentences (0–0.15 on [-1,1] range)
  - bigger spread means the system is more sensitive to the race words

Bias Learned from Training Data

We analyzed the predictions of our baseline system

- SVM model learned from the official training dataset
- features: word unigrams
- no other language resources used

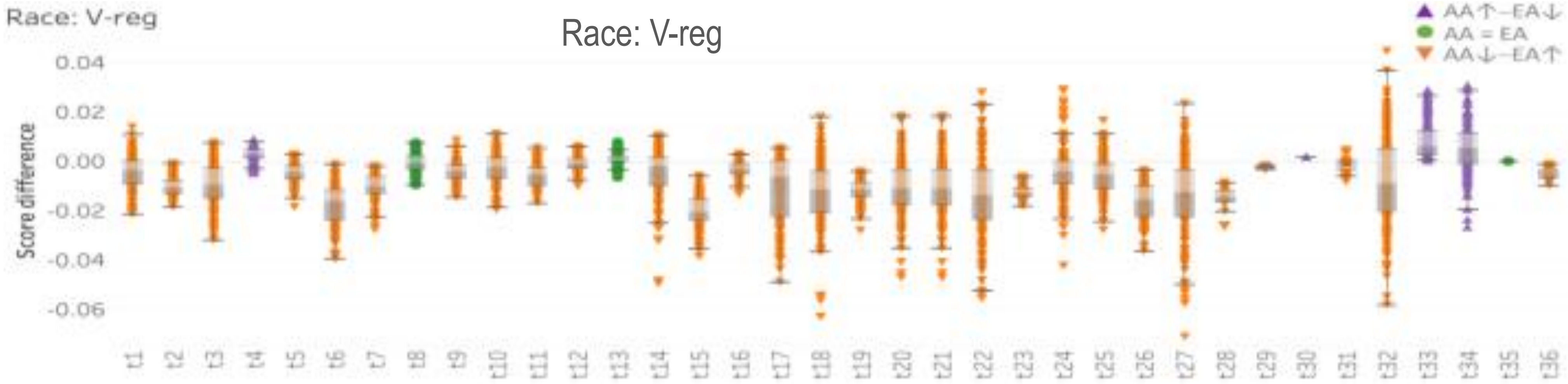


Gender bias results on the Equity Evaluation Corpus.
Teams are ordered left to right by their performance on the Tweets Test Set.

Bias Learned from Training Data

We analyzed the predictions of our baseline system

- SVM model learned from the official training dataset
- features: word unigrams
- no other language resources used



Race bias results on the Equity Evaluation Corpus.
Teams are ordered left to right by their performance on the Tweets Test Set.

baseline system

Bias Learned from Training Data

- Training data contains some biases
- The training datasets were created using a fairly standard approach:
  - polling Twitter with task-related query terms (in this case, emotion words)
  - manually annotating the tweets with task-specific labels

Data collected by distant supervision can be a source of bias

Informing Participants about the Mystery Set, Bias Experiments, Results

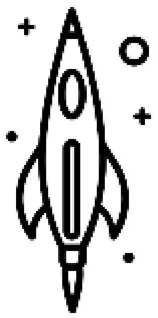
Post-Competition:

- Emailed participants the purpose of the mystery set
- Emailed each team their individual bias results
- Posted paper describing the bias experiment and aggregated results on the task website
- Presented talk about the experiment

Summary

- Created the Equity Evaluation Corpus (EEC):
  - 8,640 sentences with gender- and race-associated words
- Used the EEC to analyze the output of 219 NLP systems
  - as part of a shared task on predicting sentiment and emotion intensity
- Biases in systems:
  - 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

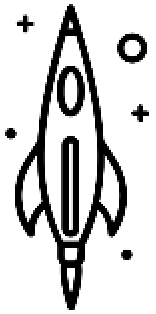
Future Work



Created by Symbolon from Noun Project

- Extend EEC by adding sentences associated with:
  - country names
  - professions (doctors, police officers, janitors, teachers, etc.)
  - fields of study (arts vs. sciences)
  - other races (Asian, mixed, etc.)
  - other genders (agender, androgyne, trans, queer, etc.)
- Identify the extent of bias in:
  - word embeddings, sentiment lexicons, lexical semantic resources, etc.
  - what is the source of bias in those resources?
- Identify the extent to which different machine learning architectures accentuate or mitigate inappropriate biases

Future Work



- Detecting bias in NLP systems

  Rudinger et al. 2018; Zhao et al. 2018

- What methods can be used to minimize biases without hurting predictive power?

  Schmidt, 2015; Bolukbasi et al., 2016; Kilbertus et al., 2017; Ryu et al., 2017; Speer, 2017; Zhang et al., 2018

- How does the quality of predictions vary when applied to text produced by different demographic groups?

  Hovy, 2015; Blodgett et al., 2016, Jurgens et al., 2017; Buolamwini and Gebru, 2018

- How to build systems that not only assign affect scores but also explain their decisions and biases?

Resources Available:

- The Equity Evaluation Corpus
  https://svkir.com/resources.html#EEC
- SemEval-2018 Task 1: Affect in Tweets
  https://competitions.codalab.org/competitions/17751

Svetlana Kiritchenko

svetlana.kiritchenko@nrc-cnrc.gc.ca