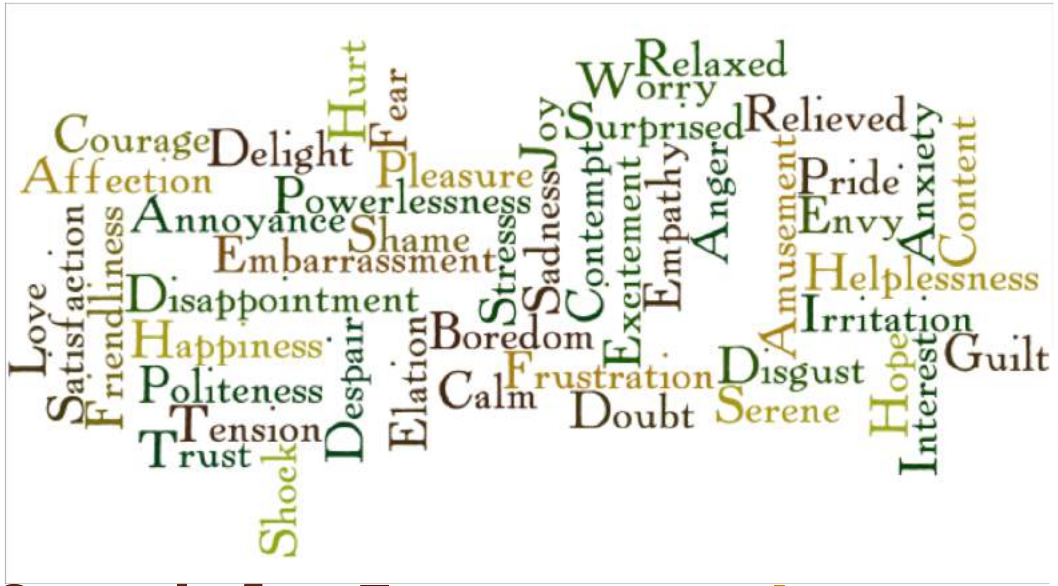
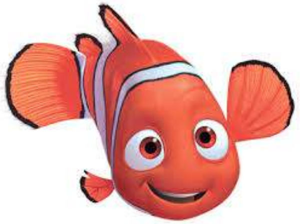
Examining Fairness in Language Through Emotions

Saif M. Mohammad
Senior Research Scientist, National Research Council Canada
Saif.Mohammad@nrc-cnrc.gc.ca @SaifMMohammad

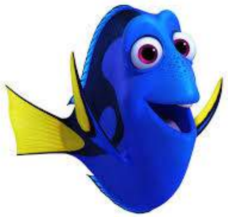
Emotions

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





The Search for Emotions in Language



creativity

fairness

SemEval-2018 Task 1: Affect in Tweets

https://competitions.codalab.org/competitions/17751

Felipe José Bravo Márquez

Five Tasks: Inferring likely affectual state of the tweeter

Mohammad Salameh

English, Arabic, and Spanish Tweets

75 Team (~250 systems)

fairness

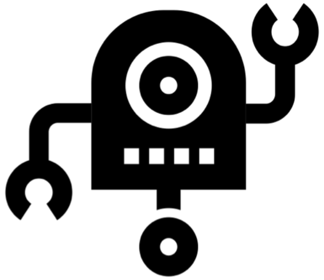
Svetlana Kiritchenko

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

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.

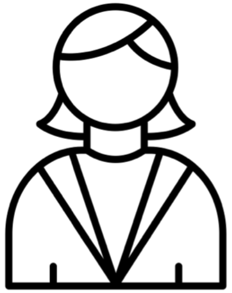
Do Machines Make Fair Decisions?

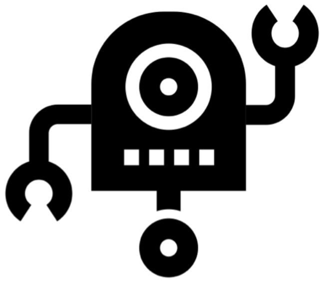
Not always—recent studies have demonstrated that as the models have become more sophisticated, they have inadvertently inherited inappropriate human biases



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



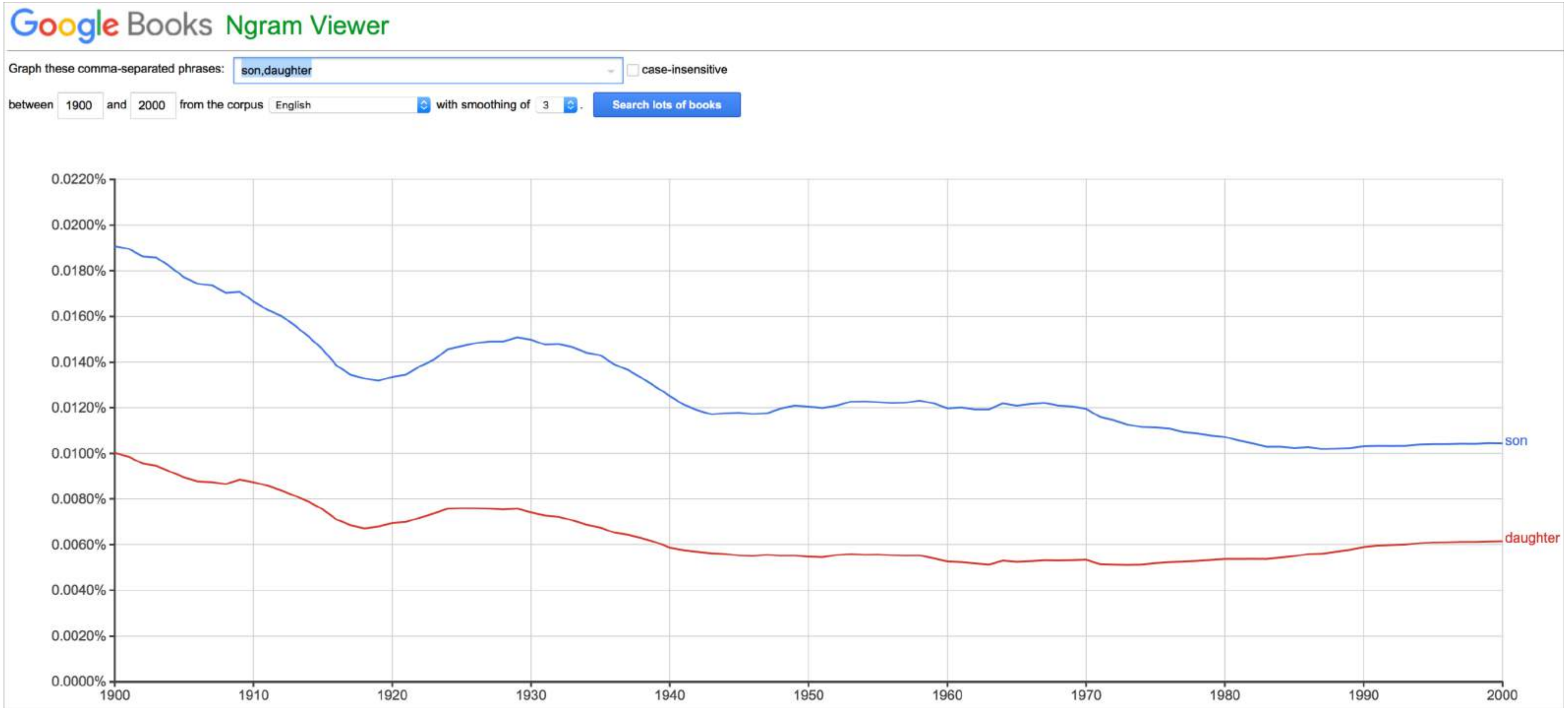


built on human data

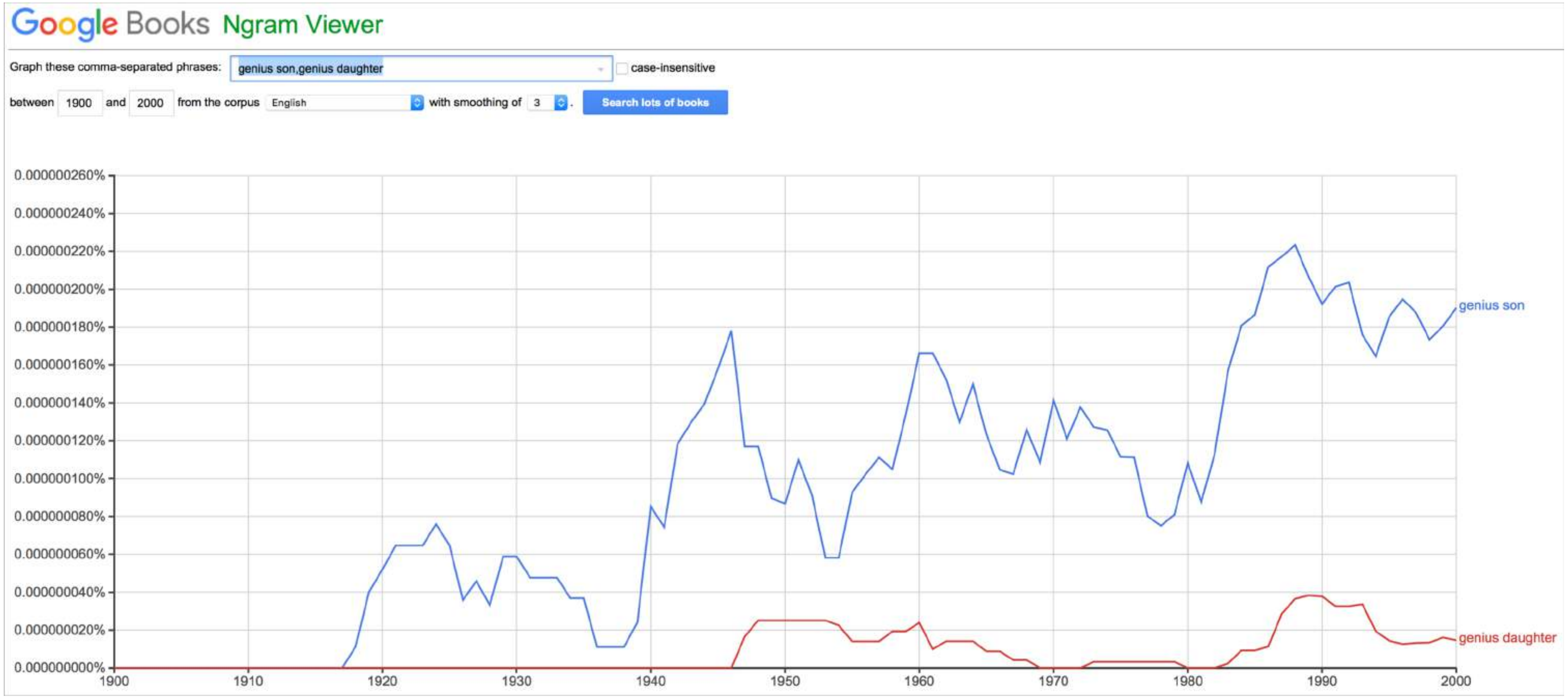
Examples of Biased AI

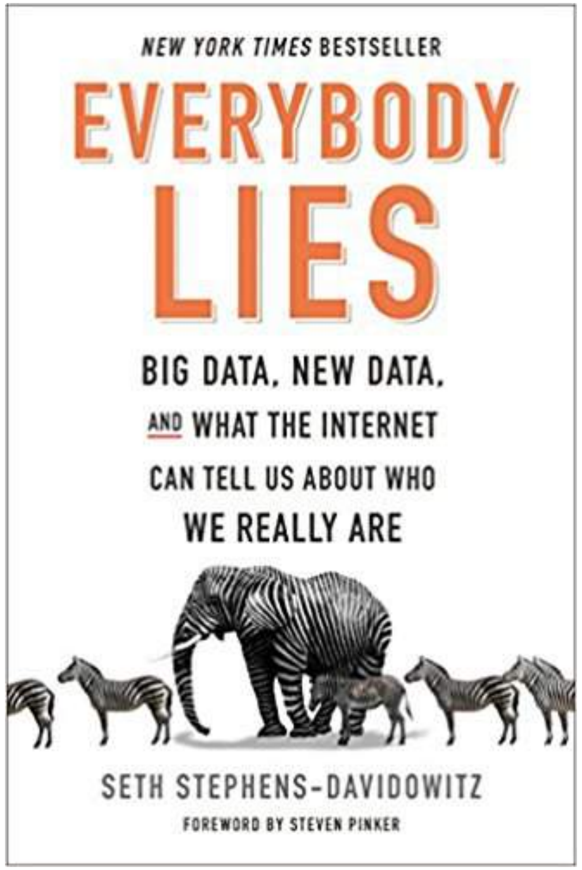
- Tay, Microsoft's racist chat bot posting inflammatory and offensive tweets
- Amazon's AI recruiting tool biased against women
  - penalized resumes that included the word "women's," as in "women's chess club captain."
- Face recognition systems good for detecting faces of white men, but really bad 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?

SemEval-2018 Task 1: Affect in Tweets

https://competitions.codalab.org/competitions/17751

Five Tasks: Inferring likely affectual state of the tweeter

English, Arabic, and Spanish Tweets

75 Team (~250 systems)

fairness

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

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.

Svetlana Kiritchenko

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems

- 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

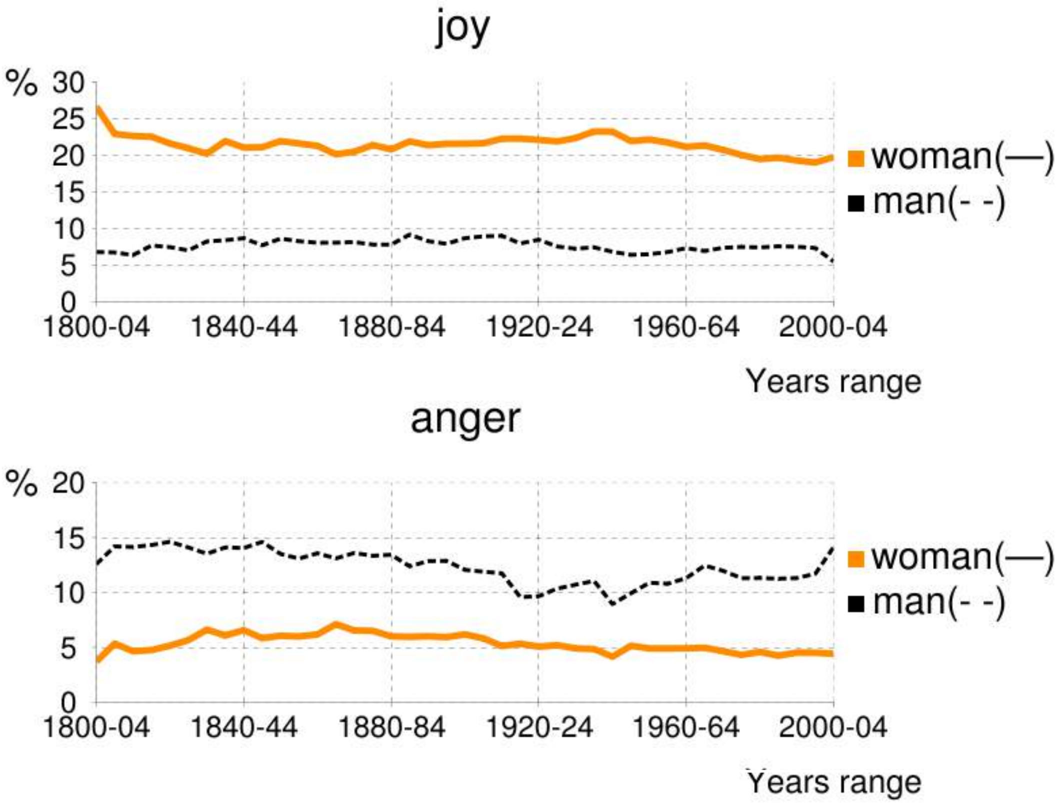
Bias Results

- 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

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.

Examining Biases in Mentions of Men and Women

- Are mentions of men and women surrounded by significantly different emotional language?



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

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

Examining Biases in Dyadic Interactions Between Men and Women

Hadrien Van Lierde

Are mentions of:

- women by women
- women by men
- men by men
- women by men

…surrounded by significantly different emotional language?

English Tweets Corpus (1B Tweets)

Google Books English Fiction Corpus (90B words)

What about work that can be deliberately abused?

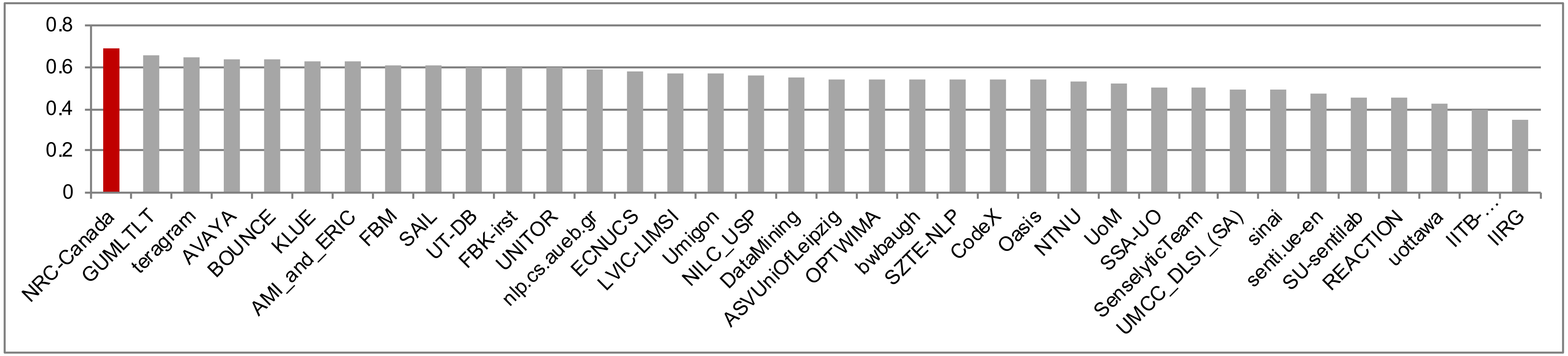
NRC at Sentiment Analysis Competitions 2013,2014

Svetlana Kiritchenko

Xiaodan Zhu

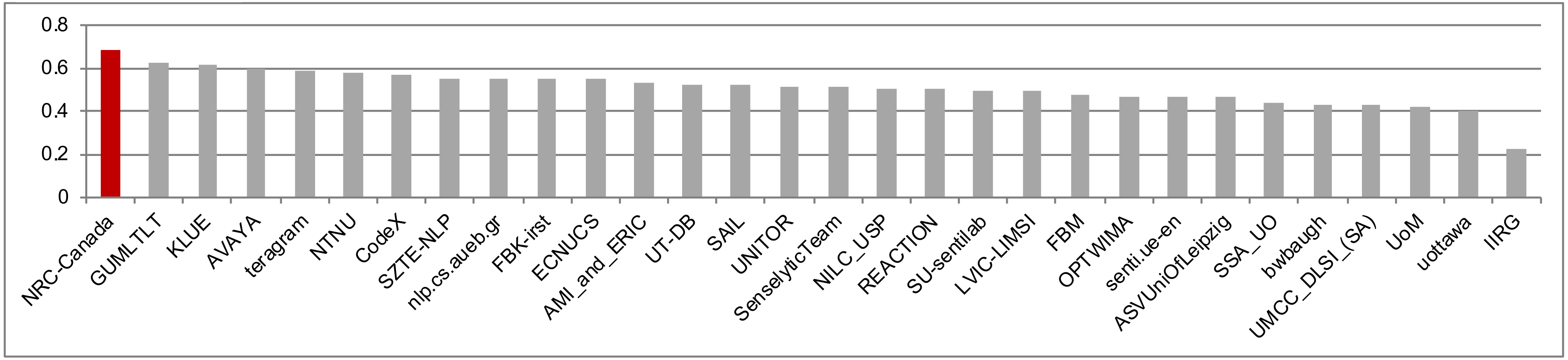
Classify Tweets: 44 teams

F-score



Classify SMS messages: 40 teams

F-score



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:

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


Xiaodan Zhu

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


Colin Cherry

Detecting Stance in Tweets

favor against neither


Parinaz Sobhani


Svetlana Kiritchenko


Xiaodan Zhu


Colin Cherry

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

But…

It is a dangerous world when computers can identify your stance on all kinds of issues. It opens the door to abuse and manipulation.

Equity does not imply sameness

Shared Understanding of VAD:
Within and Across Demographic Groups

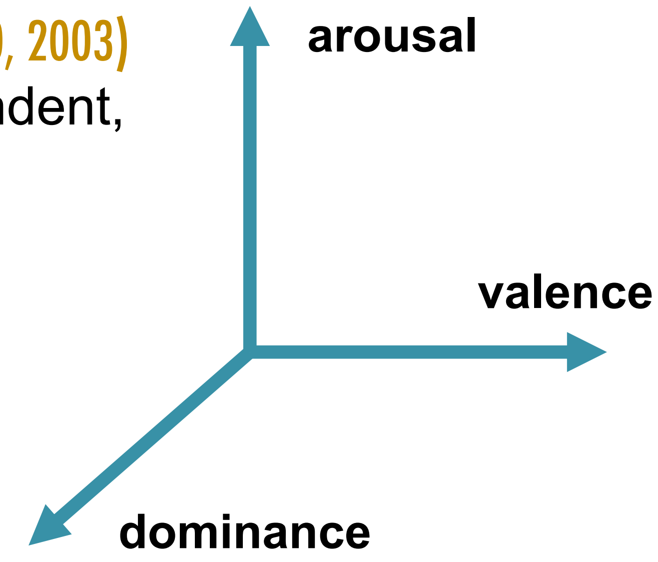

fairness

- 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

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*

fine-grained

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance

for 20,000 English Words

used comparative annotations (and not rating scales)

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.

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).

Substantially More Reliable than Past Lexicons

High Split-Half Reliability Scores (>0.9)

Research Question: Do different demographic groups differ in how they rank words by V, A, and D?

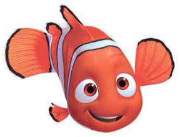
Analysis of VAD Judgments by Different Demographic Groups

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

This raises further questions:

- why do these differences exist?
- to what extent should these differences exist?

Summary  

Machine learning systems that learn from human data have inappropriate biases

We need work on:

- Measuring inappropriate biases in AI systems and inappropriate biases in language
- Developing algorithms to prevent and mitigate inappropriate biases

Machine learning systems can be intentionally used to harm and manipulate

We need work on:

- How to avoid and mitigate abuse

Equity does not imply sameness

We need work on:

- Measuring and tracking commonalities and differences across demographic groups

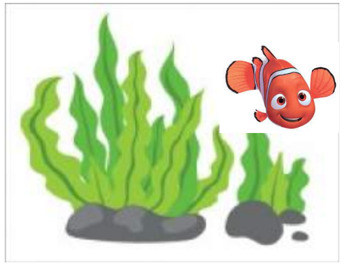
Resources Available at: www.saifmohammad.com

- 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



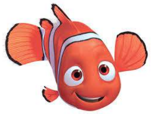
Pictures Attribution

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

Two Parts To The Work

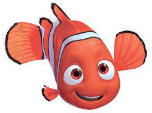
The Search for Emotions – by Humans



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.