Quantifying Gender Bias in Different Corpora

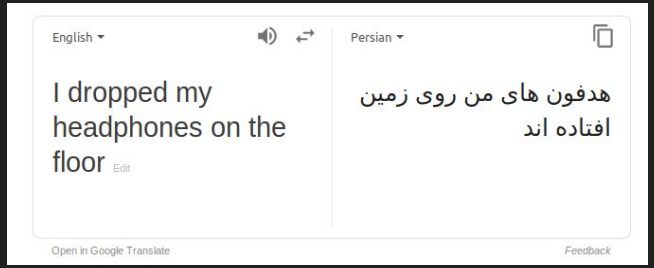
Marzieh Babaeianjelodar
Josh Gordon
Jeanna Matthews

Introduction

- A part of my thesis research
- A group of researchers
- Accepted in FATES on the Web Conference

Machine Learning Systems

- Machine Learning ?
  - Train examples
  - Learns like a human (e.g. Amazon Alexa)
  - Can predict the future
  - Examples: Machine Translation, Sentiment Analysis
  - Computer Vision (train images)



Effect of NLP in our lives

- Natural Language Processing: NLP
- NLP: Making a computer understand human language
- Used in different applications:
  - Resume parsing
  - Predicting the oscars (based on movie ranks)
  - Computer Science students (applications for universities)
- NLP systems can carry Bias:
  - Through the algorithm
  - Through the data
  - Hiring more male students rather than female students

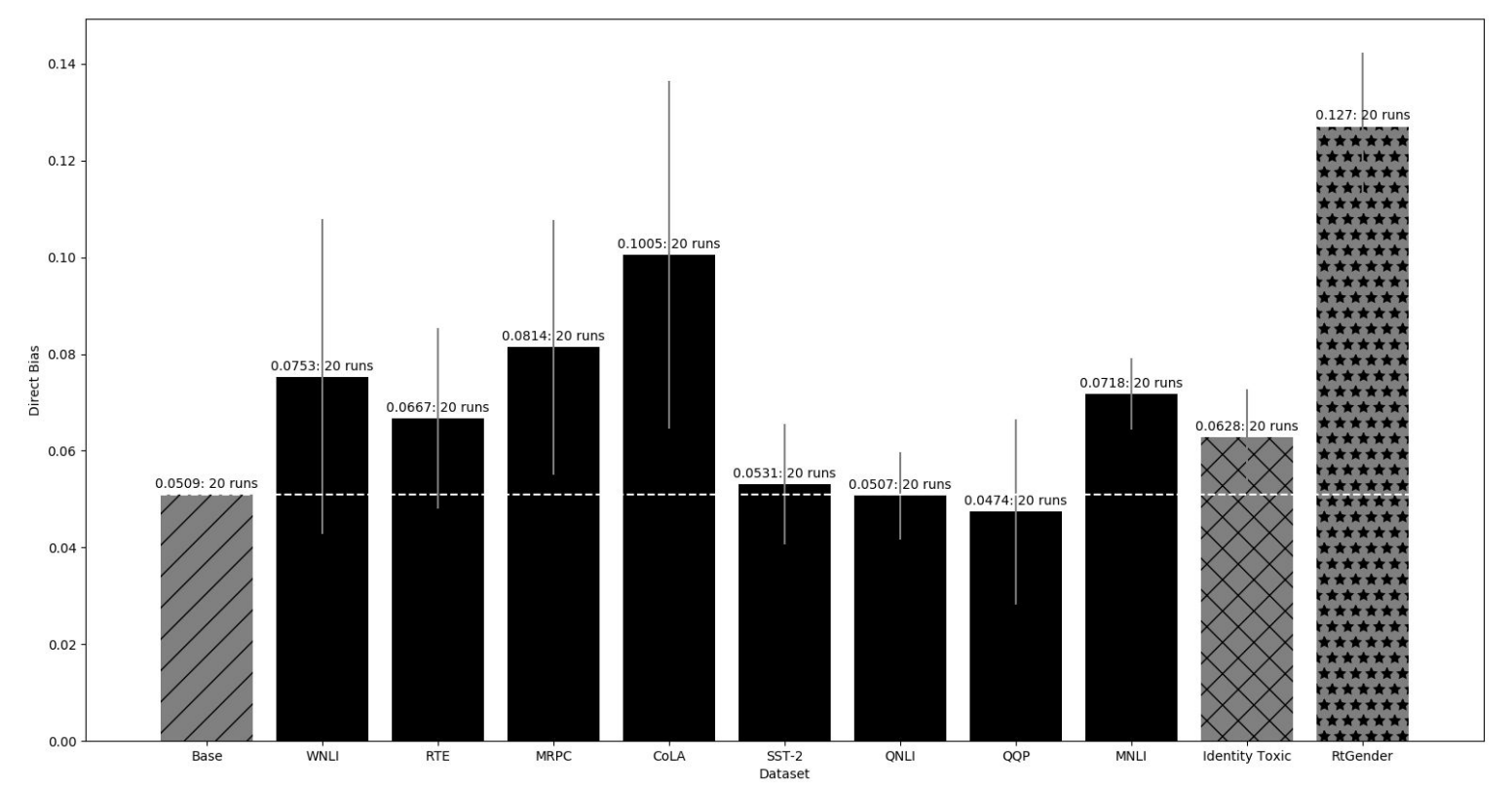
Data

- Body of text to feed to the NLP systems:
  - GLUE: General Language Understanding Evaluation
  - RTGender: labeled based on (source of response to social media comments) f and m
  - Toxic identity: labeled based on attacking race, gender, religion, homosexuality, etc
- Importance: How the system carries these biases

Word Embeddings and Gender Bias

- Words are represented as positions in space
- The distribution is learned from the data
  - Position carries semantic meaning
  - Position can encode gender
- Can compute “gender direction”
- Bias is represented by average similarity to this direction

Our Results


0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Direct Bias
0.0509: 20 runs
0.0753: 20 runs
0.0667: 20 runs
0.0814: 20 runs
0.1005: 20 runs
0.0531: 20 runs
0.0507: 20 runs
0.0474: 20 runs
0.0718: 20 runs
0.0628: 20 runs
0.127: 20 runs
Base
WNLI
RTE
MRPC
CoLA
SST-2
QNLI
QQP
MNLI
Identity Toxic
RtGender
Dataset

Importance of Thorough Testing

- Biases show up in seemingly innocuous systems
- Can be unclear who is accountable
  - Wider issues with giving computers agency
- Debiasing algorithms

Thank You for your Attention