Analyzing the Effects of Annotator Gender Across NLP Tasks

Laura Biester✧, Vanita Sharma✧, Ashkan Kazemi✧, Naihao Deng✧, Steve Wilson★, Rada Mihalcea✧

✧ University of Michigan, Computer Science & Engineering

★ Oakland University, Computer Science & Engineering

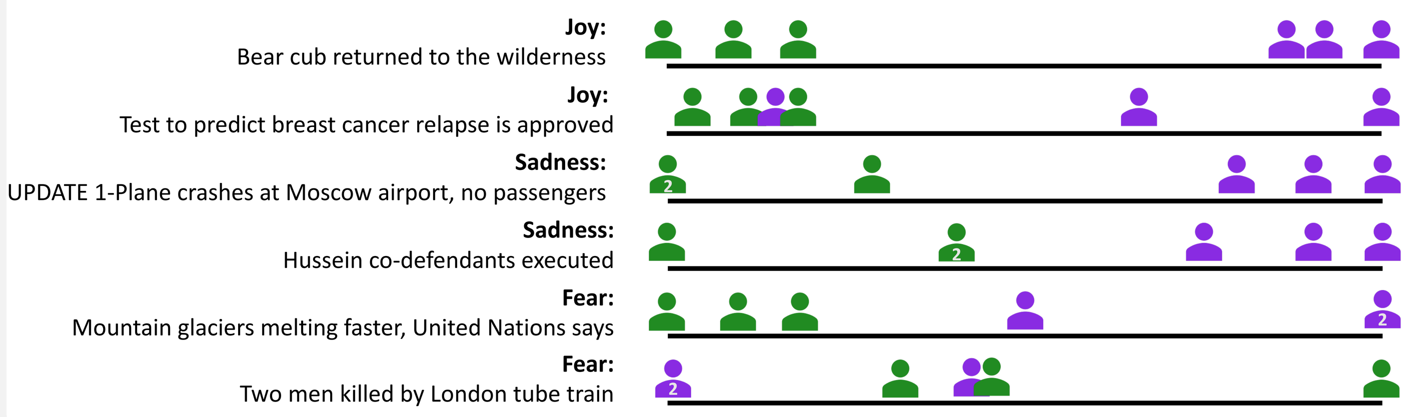
Code, Paper, Data

contact: lbiester@umich.edu

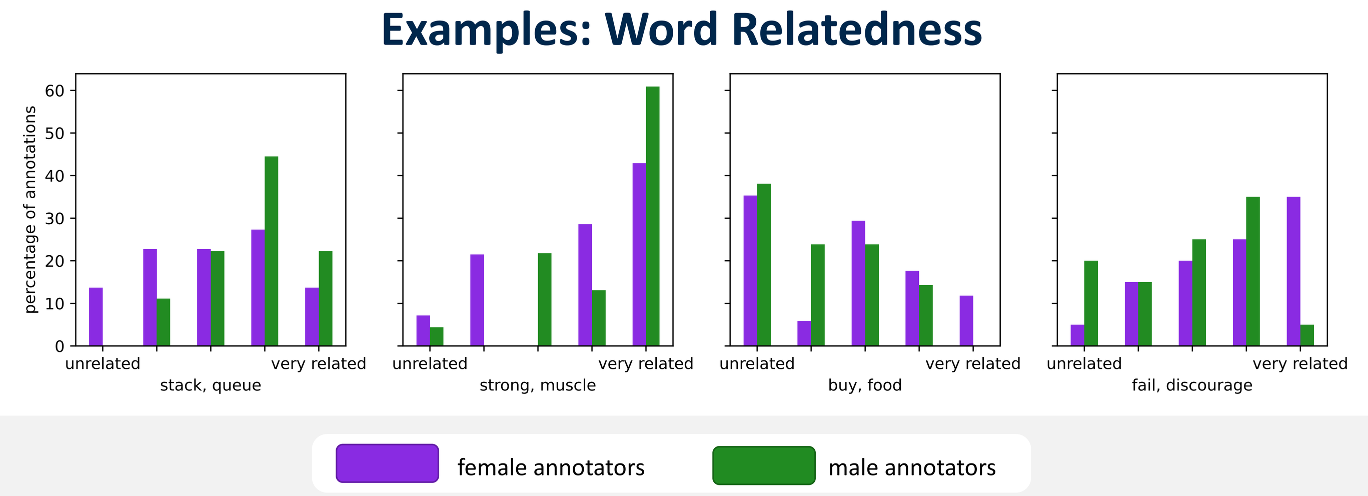
Motivation

- Work in areas such as hate speech detection has revealed clear differences in annotation based on the demographic groups of annotators
- We do not know how much annotator demographics affect a broader range of NLP tasks
- Annotator differences can cause problems with generalization to new users

Examples: Affective Text



Examples: Word Relatedness



Methodology

Distribution Analysis

- Plot distributions of scores
- Run permutation tests with random gender assignments to determine significance
- *Interval data:* is the area between two distribution curves different?
- *Ordinal data:* are visible differences in the distribution significant?

Agreement Analysis

- Agreement computed with Krippendorff's Alpha
- Compute agreement scores between each annotator and aggregate of other annotators
  - *All, same gender, different gender annotators*
- Plot results and run t-tests for interesting pairs

Data

Identifying Data

- Surveyed NLP papers to see if they collected annotator demographics (most did not mention demographics)
- Emailed authors of 23 datasets, and most authors who replied stated that they did not collect demographics
- The datasets we used were chosen due to ***accessibility***, but still cover an interesting ***variety of tasks***
- Due to data size, limited to binary gender

Dataset	Male Annotators	Female Annotators	Datapoints	Annotations per Datapoint (mean)	Annotation Type	Ratings per Datapoint
Affective Text	3	3	1000	6.00	Interval	7
Word Similarity	196	157	498	38.74	Ordinal	2
Sentiment Analysis	736	744	14071	4.21	Ordinal	1
NLI	282	211	1200	9.26	Ordinal	1

Affective Text

- SemEval 2007 Task 14 (Strapparava and Mihalcea, 2007)
  - Disaggregated labels released with our paper
- Six emotions (anger, disgust, fear, joy, sadness, and surprise) + valence
- Emotions 0—100, valence -100—100

Word Similarity

- Word pairs rated for similarity and relatedness on a 5-point Likert scale
- 25% of pairs from SimLex-999 (Hill et al., 2015)
- 75% of pairs inspired by Garimella et al. (2017)
  - Pairs chosen due to discrepancies in Indian, US, male, and female word associations

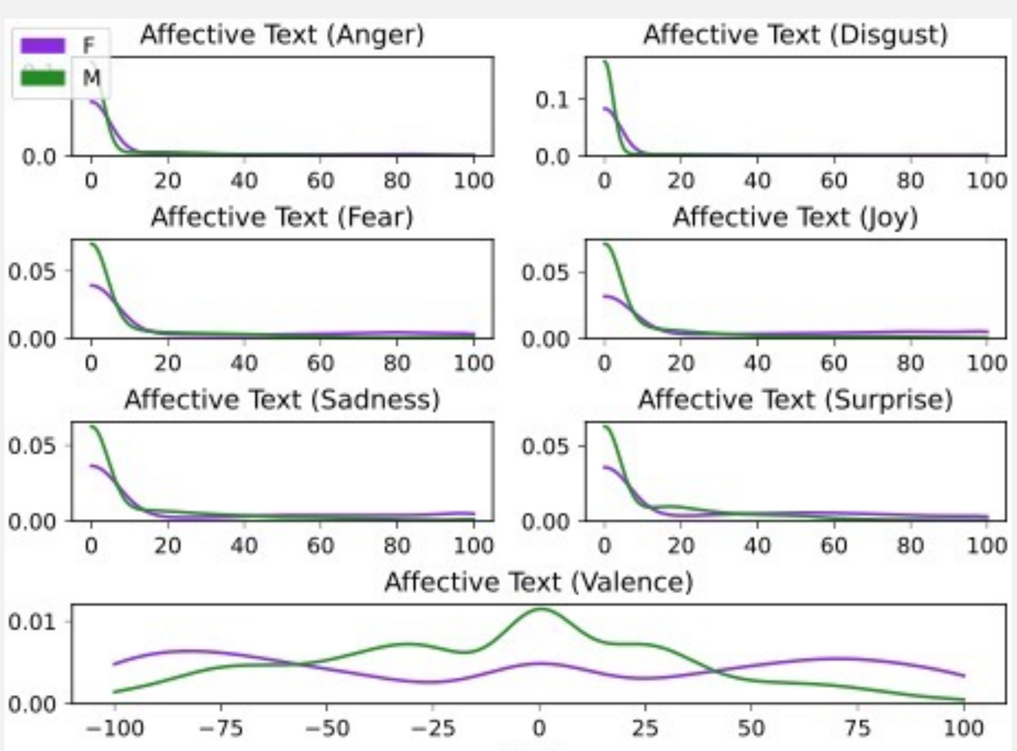
Sentiment Analysis

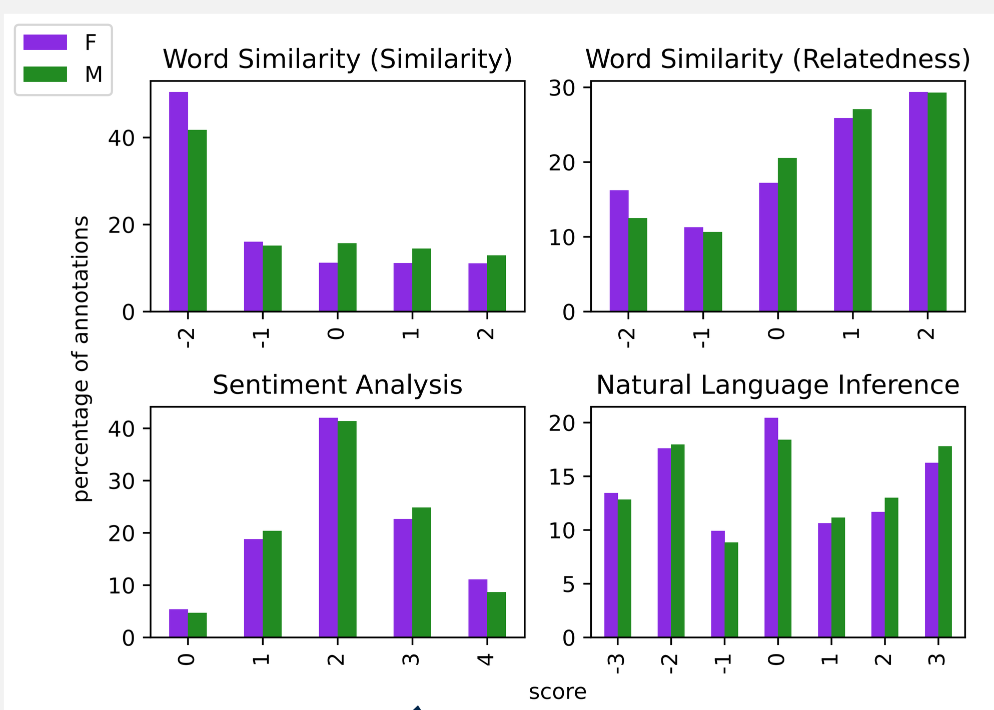
- Dataset for measurement of age-related bias in sentiment analysis (Diaz et al., 2018)
- Training data text drawn from Sentiment140 (Go et al., 2009)
- 5-point Likert scale (very negative — very positive)

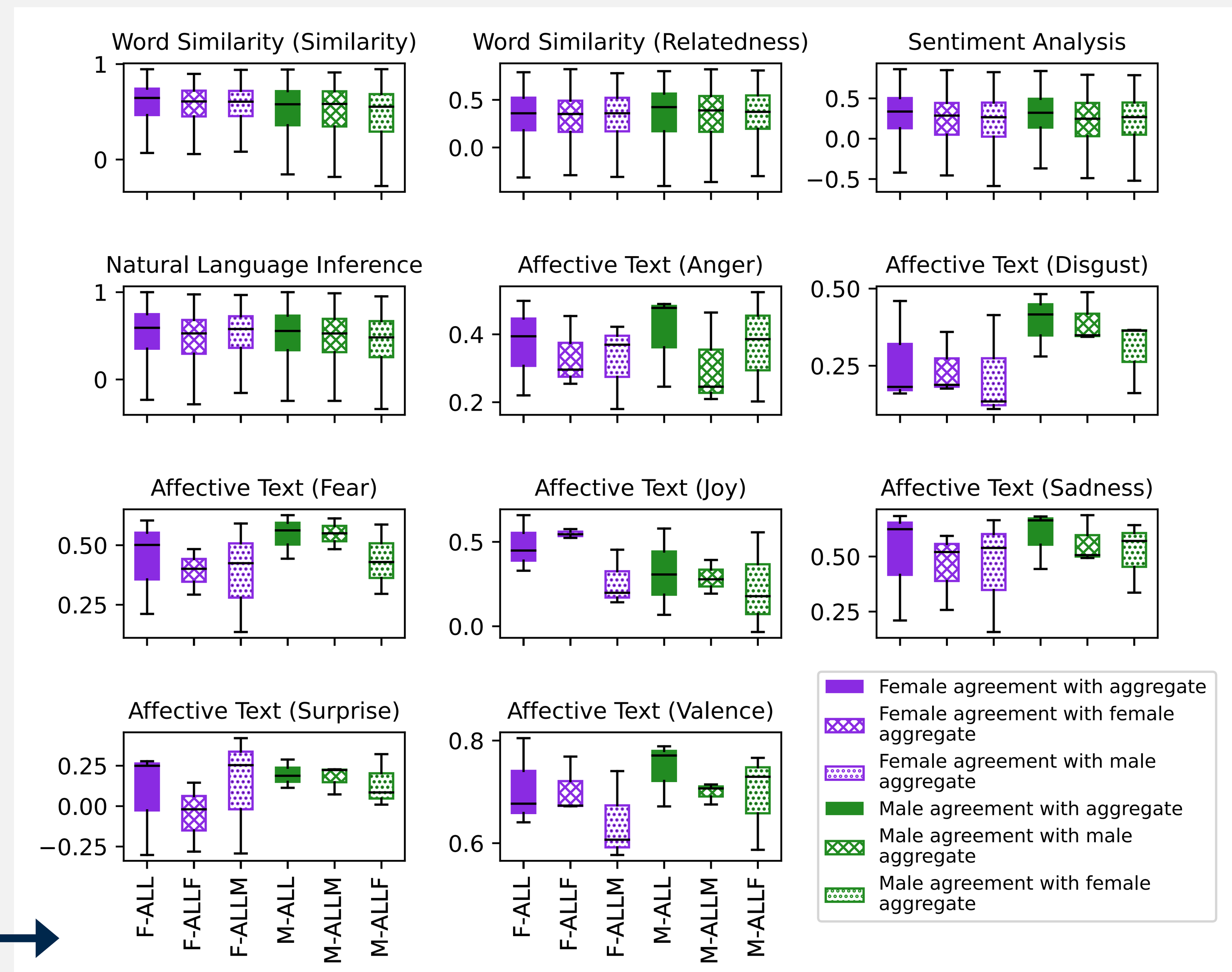
Natural Language Inference (NLI)

- CommitmentBank (De Marneffe et al., 2019)
  - *Demographics provided by the author, not publicly available*
- 7-point Likert scale: does the annotator believes that the author of the text is certain that the prompt is true or false?

Results







- **Significant difference** for sentiment analysis distribution
  - Men give more intermediary labels (somewhat positive/somewhat negative)
- Other visible patterns (e.g., higher word similarity scores for men) not significant according to permutation testing
- *No significant differences* found in agreement comparisons