ITCS 4111/5111: Introduction to NLP

Biases vs. Fairness and Rationality

Razvan C. Bunescu

Department of Computer Science @ CCI

rbunescu@uncc.edu

Bias and Fairness in NLP Models

- Harms in **sentiment classifiers**:
  - Kiritchenko and Mohammad (2018) found that several of the 200 systems evaluated consistently provide slightly higher sentiment intensity predictions for one race or one gender.
    - Example: out of 200 sentiment classifiers, most assigned lower sentiment and more negative emotion to sentences with African American names in them.
    - Effect: Perpetuate negative stereotypes.

- Harms in **toxicity classification**:
  - Toxicity detection is the task of detecting hate speech, abuse, harassment, or other kinds of toxic language.
  - Some toxicity classifiers incorrectly flag as being toxic sentences that are non-toxic but simply:
    - Mention identities like blind people, women, or gay people, or
    - Use linguistic features characteristic of varieties like African-American Vernacular English.
    - Effect: censorship of discussion about these groups.

Where does the bias come from?

- Most state-of-the-art NLP models are trained on large amounts of data.
  - ML models' behavior largely influenced by the training data:

> *The (deep learning) algorithm is simple, **the model's complexity comes from the data**.*
> - Need algorithms that can scale easily to large amounts of data.
>   - See also Richard Sutton: The bitter lesson.

- Most bias comes from the **data**:
  - The training data: ML systems are known to amplify the biases in their training data.
  - The human labels.
  - The resources used (like lexicons).

- Bias can also come from the ML **architecture**:
  - What the model is trained to optimized, i.e. the loss function.

Bias and Fairness

Hort et al., 2023 "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey"

- ML algorithms, which are not designed to intentionally incorporate bias, run the risk of **replicating** or even **amplifying bias** present in real-world data.
  - advertisement and recruitment processes, university admissions, human rights, …

- This may cause **unfair treatment** in which some individuals or groups of people are privileged (i.e., receive a favourable treatment) and others are unprivileged (i.e., receive an unfavourable treatment).

- A **fair treatment** of individuals requires that:
  - Decisions are made independent of *sensitive attributes* such as gender or race.
    - Often called **protected attributes**.
  - Individuals are treated based on **merit**.

Bias Mitigation is an Open Research Area

Hort et al., 2023 “Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey”

1. Use a **diverse** and **representative dataset**.

2. Define **protected attributes** (e.g. gender, religion) and ensure classifier (e.g. loan approval) does not depend on them.
   - Not as simple as removing protected attributes:
     - Protected attributes can be correlated with other attributes, e.g. race and religion may be correlated with zipcode, city or neighborhood.
   - Relabeling, sampling, weighting, …
   - Use regularization and constraints with the loss function.
   - Representation learning, adversarial learning.

3. Reinforcement Learning from Human Feedback (**RLHF**).

Model Cards

Mitchell et al., 2019. Model Cards for Model Reporting

- For each algorithm you release, document:
  - Training algorithms and parameters.
  - Training data sources, motivation, and preprocessing.
  - Evaluation data sources, motivation, and preprocessing.
  - Intended use and users.
  - Model performance across different demographic or other groups and environmental situations.

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

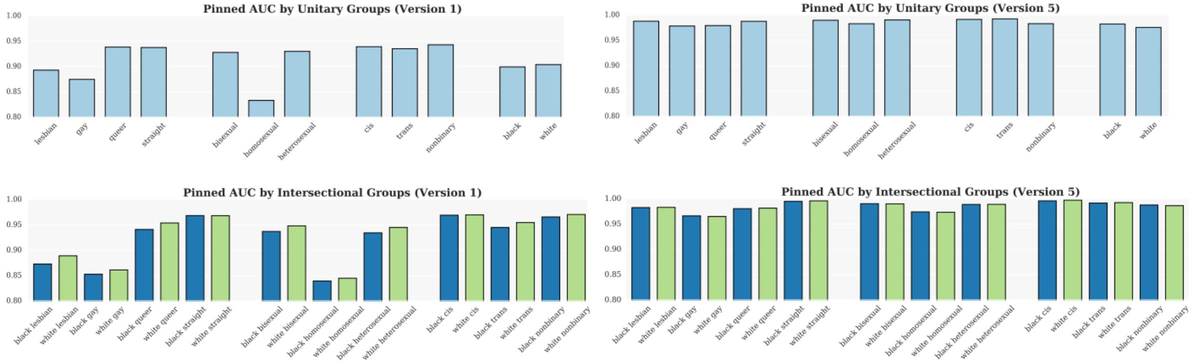
Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses



Pinned AUC by Unitary Groups (Version 1)

Pinned AUC by Unitary Groups (Version 5)

Pinned AUC by Intersectional Groups (Version 1)

Pinned AUC by Intersectional Groups (Version 5)

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

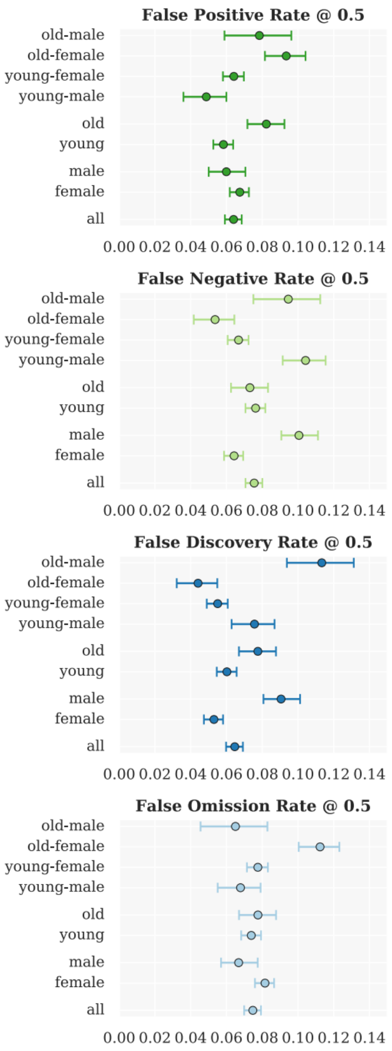
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



False Positive Rate @ 0.5

False Negative Rate @ 0.5

False Discovery Rate @ 0.5

False Omission Rate @ 0.5

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Cognitive Biases

- Over 180 documented **cognitive biases** that pervade human reasoning and decision making that are routinely ignored when discussing the ethical complexities of AI.
  - **Bounded rationality** (Herbert Simon, 1957): our mental capacity for making fully rational decisions is influenced by limitations in human cognition and one's environment.
    - *To become an expert on a topic requires about ten years of experience.* (also Norvig …)
    - These limitations result in the use of *heuristics*, or mental shortcuts, that help individuals reason and make decisions using simple, yet typically effective, strategies.
    - They extend from social interaction to judgment and decision making.
  - When cognitive bias is present, **faulty reasoning**, **irrationality**, and potentially **detrimental outcomes** (e.g., financial losses, health disparities, environmental impact) can result.

- Language Models (LM) are susceptible to cognitive biases:
  - Base rate neglect and value selection bias, Anchoring and adjustment, Framing effects, …

Talboy and Fuller, 2023: Challenging the appearance of machine intelligence: Cognitive bias in LLMs and best practices for workplace adoption

Representativeness

- Tversky & Kahneman, 1974:
  - Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail. Order the probability of Steve being in each of the following occupations:
    - Farmer
    - Salesman
    - Airline pilot
    - Librarian
    - Middle school teacher

  - Problem asks for P(Steve's job is X | Steve is shy …)
    - This is different from P(Steve is shy … | Steve's job is X)

Insensitivity to sample size

- Tversky & Kahneman, 1974:
  - A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.
  - For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?
    - The larger hospital.
    - The smaller hospital.
    - About the same.

Base rate neglect

- Bar-Hillel, 1978:
  - 10 out of every 1,000 women at age forty who participate in routine screen have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography.
  - Here is a new representative sample of women at age forty who got a positive mammography in routine screening. What is their likelihood of having breast cancer?

Hint: "Real truth seeking is Bayesian".

Anchoring

- Tversky & Kahneman, 1974:
  - A wheel of fortune with numbers between 0 - 100 was recently spun and landed on 95. Indicate whether this number is higher or lower than the percentage of African countries in the United Nations.

Proposed Best Practices for LLM Use in the Workplace

1. If LLMs are adopted, they should be used as decision support tools, not final decision makers.
2. Individual users are accountable for their LLM use, as well as the intended and unintended consequences of that use.
3. Adherence to regulations and laws regarding fair and nondiscriminatory use of technology must be upheld.

Talboy and Fuller, 2023: Challenging the appearance of machine intelligence: Cognitive bias in LLMs and best practices for workplace adoption