Detecting bias towards people in fake news classifiers using explainability methods

Dawid Płudowski, Antoni Zajko, Mikołaj Roguski, Piotr Robak

Warsaw University of Technology

December 11, 2024

Motivation

Automated Fake news detection

- Many AI fake news detectors are proposed each year
- These algorithms have a growing control over what may be published on the internet

Explainability and Fairness

- Bias in the models may infringe the right to free speech
- Bias towards specific persons is not widely studied

Research Question:

- Is model X biased toward person Y?

Aims

Leveraging biased models

- Show how bias can be used to misuse the model

Bias quantification

- Calculate bias towards specific people – how easy is it to create fake news about certain people that will not be detected?

Mitigation

- Propose measures to improve model fairness – how can we prevent misusing bias in models?

What do we use?

Data:

- **LIAR**
- COAID
- ISOT

Models:

- **RoBerTa**
- KnowBert
- Gemini (?)

How do we explain?

Attribution. Methods to assign importance to each element of the input. for this purpose, we use feature ablation which is suitable for black-boxes.

Counterfactual. Methods to introduce minimal changes to the input that result in different model predictions. We use our **custom** approach.

Are person-related tokens important?

Table: Table containing basic statistics about datasets. From the top: number of observations, average observation text length, average number of ners in an observation, average ratio of NERs to text length (in tokens) and ratio of fake and factual news.

Dataset	coaid	isot	LIAR
Observations	5457	44954	12796
Avg. text len.	66.5	80.1	107.1
Avg. # NER	0.668	1.15	0.78
# NER / Text len	0.058	0.076	0.037
Fake / True	0.17	0.48	0.47

Are person-related tokens important?

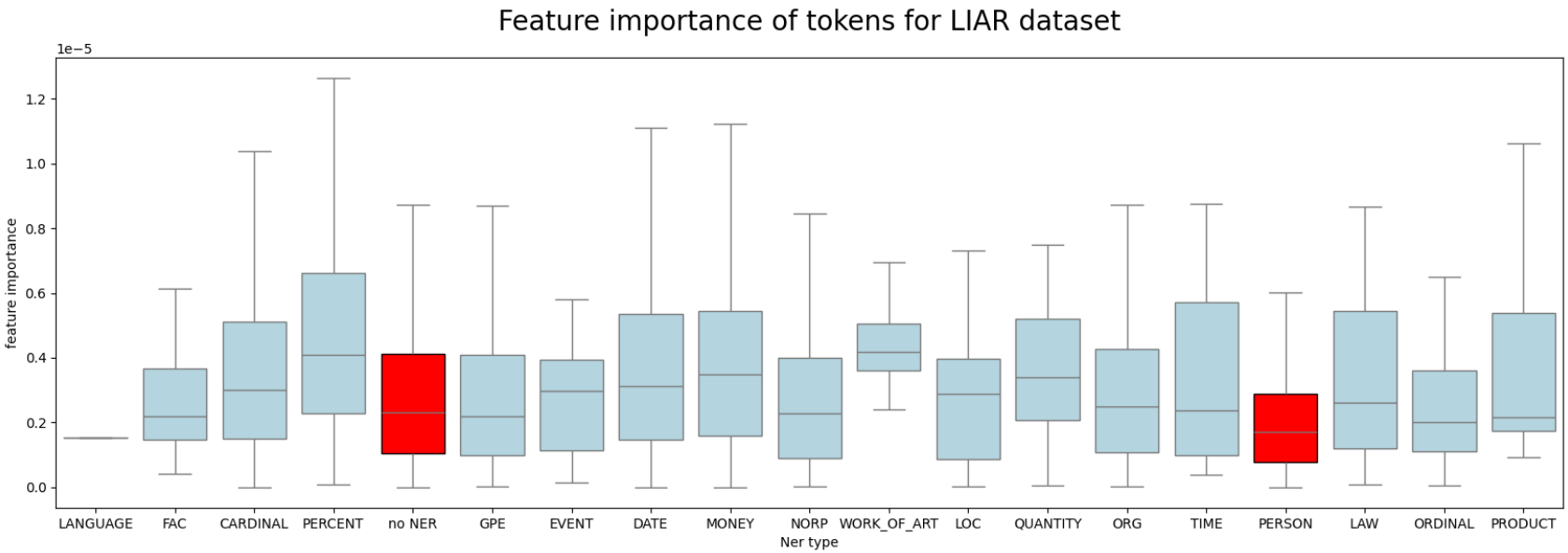

Figure: Feature importance of each NER,

How can bias be used?

Example

Mitt **Romney** *drove to Canada with the family dog Seamus strapped to the roof of the car.* – 8% probability of fake news.

Example

Mitt **Obama** *drove to Canada with the family dog Seamus strapped to the roof of the car.* – 79% probability of fake news.

How can bias be used?

Example

Toomey and **Trump** *will ban abortion and punish women who have them.* – 7% probability of fake news.

Example

Toomey and **Obama** *will ban abortion and punish women who have them.* – 68% probability of fake news.

Mitigation measures

Dataset	**Accuracy**
LIAR	0.655 +/- 0.006
LIAR without persons	0.664 +/- 0.015
COAID	0.979 +/- 0.005
COAID without persons	0.982 +/- 0.002
ISOT	0.841 +/- 0.225
ISOT without persons	0.935 +/- 0.049

Table: Comparison of accuracies of models trained on datasets with and without persons.

Challenges

- **Fine-tuning of the models** – several hours of the local machine.
- **Mapping of tokens** – NERs and models' tokens are represented differently.
- **Constructing counterfactual methodology** – how do we ensure appropriate swapping? How to handle names and surnames differently?

Future works

- Quantifying bias.
- Adding LLM to the benchmark.
- Automation of existing analysis code.
- Verify the potential reasons for the model's bias.

Thank You for attention!