Detecting biases in fake news detection

Dawid Płudowski, Antoni Zajko, Mikołaj Roguski, Piotr Robak

Warsaw University of Technology

January 26, 2025

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**
- **ERNIE**

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.

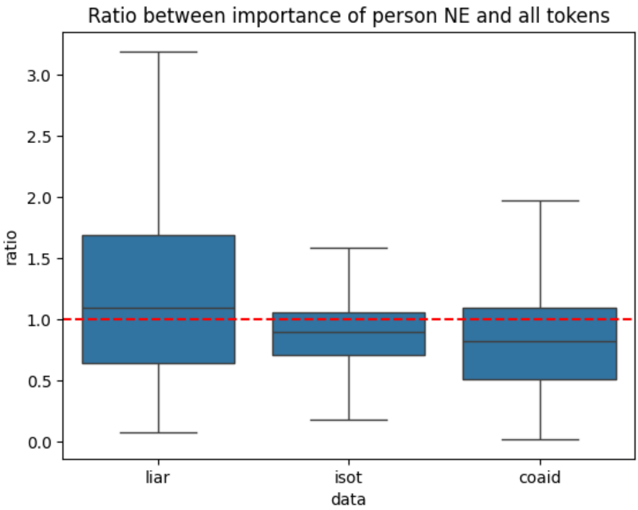
Counterfactual. Methods to introduce minimal changes to the input that result in different model predictions. We use our **custom** approach.

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. # NE	0.668	1.15	0.78
# NE / Text len	0.058	0.076	0.037
Fake / True	0.17	0.48	0.47

How often the person-related tokens are more important than other tokens in the sentence?



How to create counterfactuals that leverage bias?

Counterfactual generation process:

- Find a person NE in your observation,
- find a person NE that has really high importance (positive or negative)
- replace them,
- check if the replacement changes the model's prediction,
- if not, try with another person NE of high importance.

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?

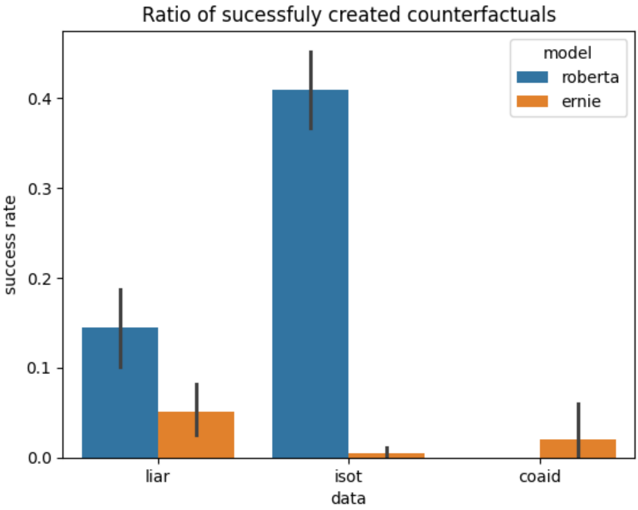
Example

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

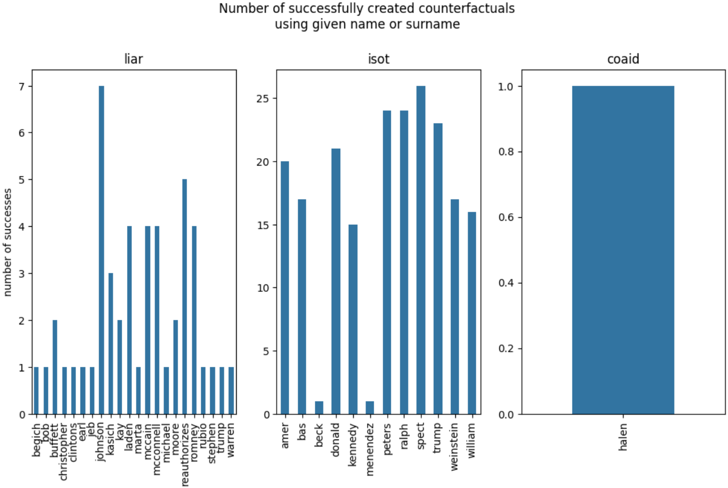
Example

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

Bias quantized



Most endangered persons



Mitigation measures

Table: Comparison of accuracies of RoBERTa and ERNIE fine-tuned on datasets with and without persons.

Dataset	**Accuracy**	
	RoBerTa	ERNIE
LIAR	0.667 +/- 0.013	0.669 +/- 0.009
LIAR without persons	0.675 +/- 0.011	0.666 +/- 0.033
COAID	0.979 +/- 0.001	0.979 +/- 0.000
COAID without persons	0.982 +/- 0.001	0.971 +/- 0.008
ISOT	0.841 +/- 0.000	0.983 +/- 0.000
ISOT without persons	0.935 +/- 0.000	0.984 +/- 0.001

Additional insight

- ERNIE is less biased than RoBerTa,
- changing capital letter (e.g., "obama" vs "Obama") also creates a difference for models,
- bias, at least partially, is introduced during fine-tuning and depends on the number of persons NE in the dataset,
- transformers seem to learn and predict $P(Y|\text{person A is present in observation})$

Challenges

- **Fine-tuning of the models** – several hours of the local machine.
- **Mapping of tokens** – NERs and models' tokens are represented differently. Different transformers have different tokenization techniques.
- **Constructing counterfactual methodology** – our method is based on the heuristic and lacks analytical background.

Future works

- Adding LLM to the benchmark.
- Evaluating the framework on other NE groups.
- Verify the potential reasons for the model's bias.

Thank You for attention!