Large scale language models are good! But are they fair?

Samia Touileb (samia.touileb@uib.no)
MediaFutures, University of Bergen

Media Futures

Norwegian datasets are mostly from news sources.

- Codified domain.

- Produced by small homogeneous samples:
  + **white**,
  + **middle-aged**,
  + **educated**,
  + **upper-middle-class**,
  + **men**.

- Trained models expect people to speak like them.

https://thumbs.dreamstime.com/b/
journalist-interviews-businessman-celebrity-reporter-holdi
jpg

How many of us actually do?
Are our models prepared to cope with reality, i.e. the **demographic variation**?

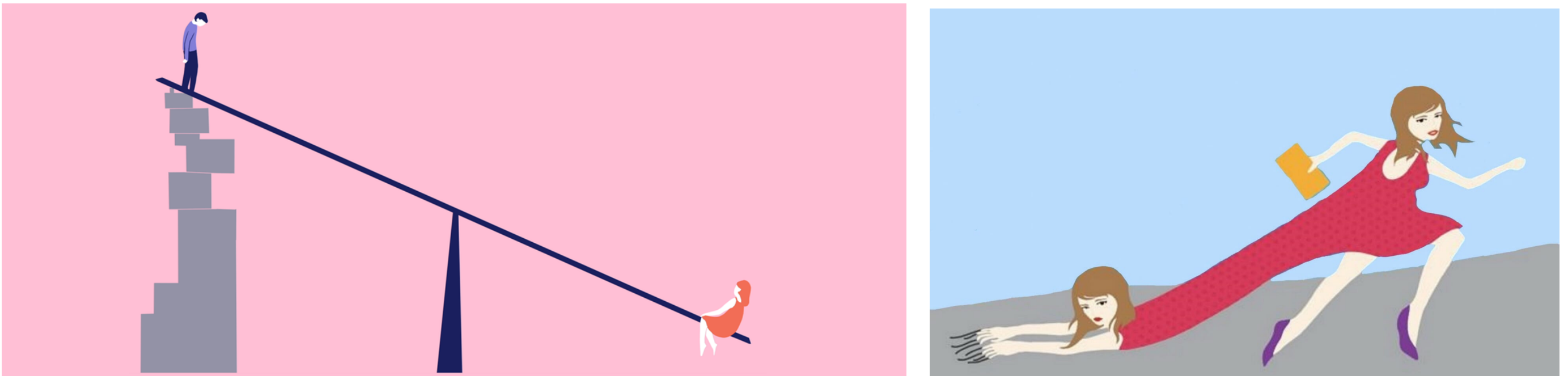
https://www.peoplemanagement.co.uk/Images/diversitygraphic_tcm27-93336.jpg

What about **gender balance?**

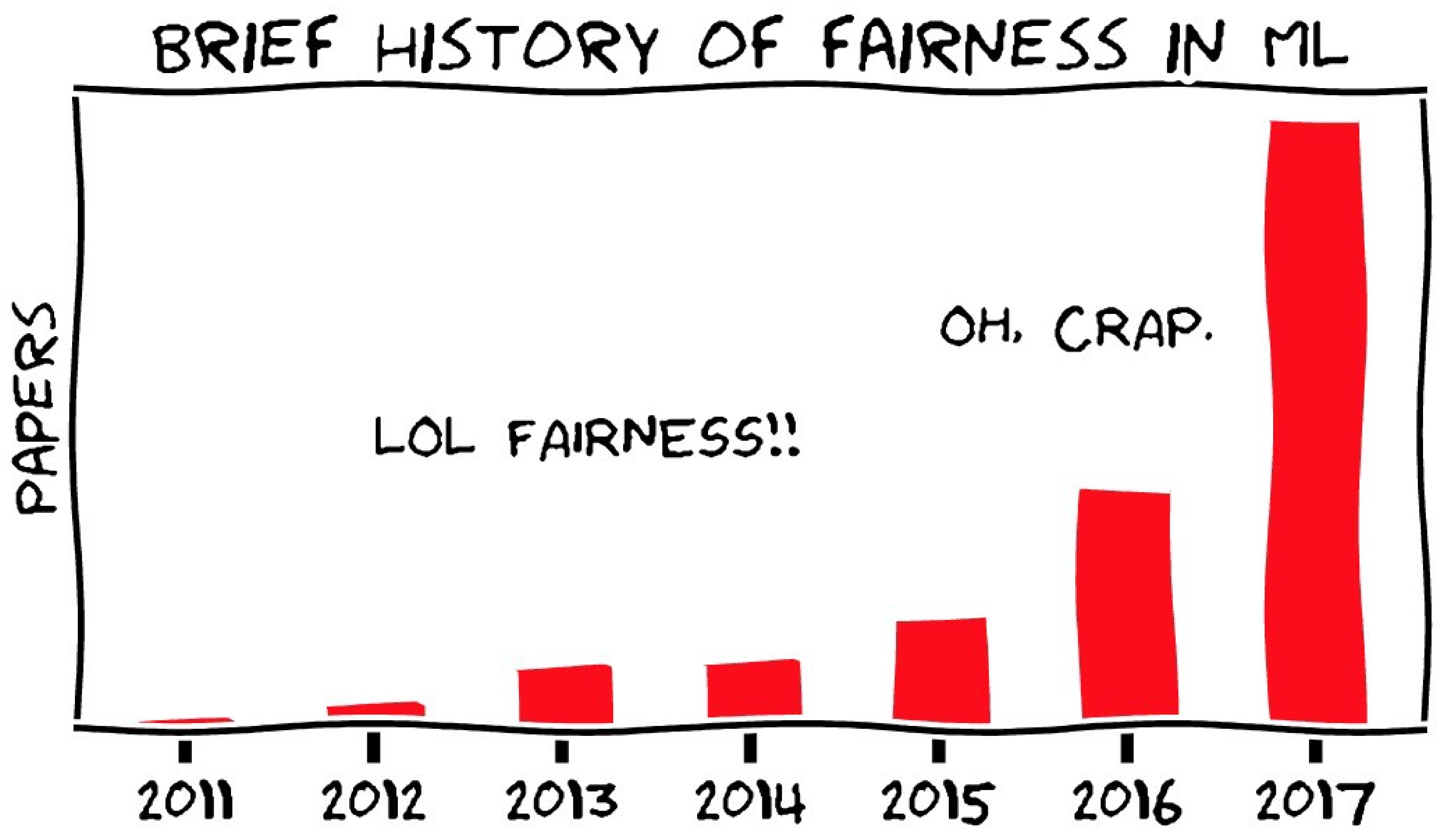
- Norwegian review data contains gender bias.

- Less books written by females are reviewed.

- Female critics are more critical of female work.



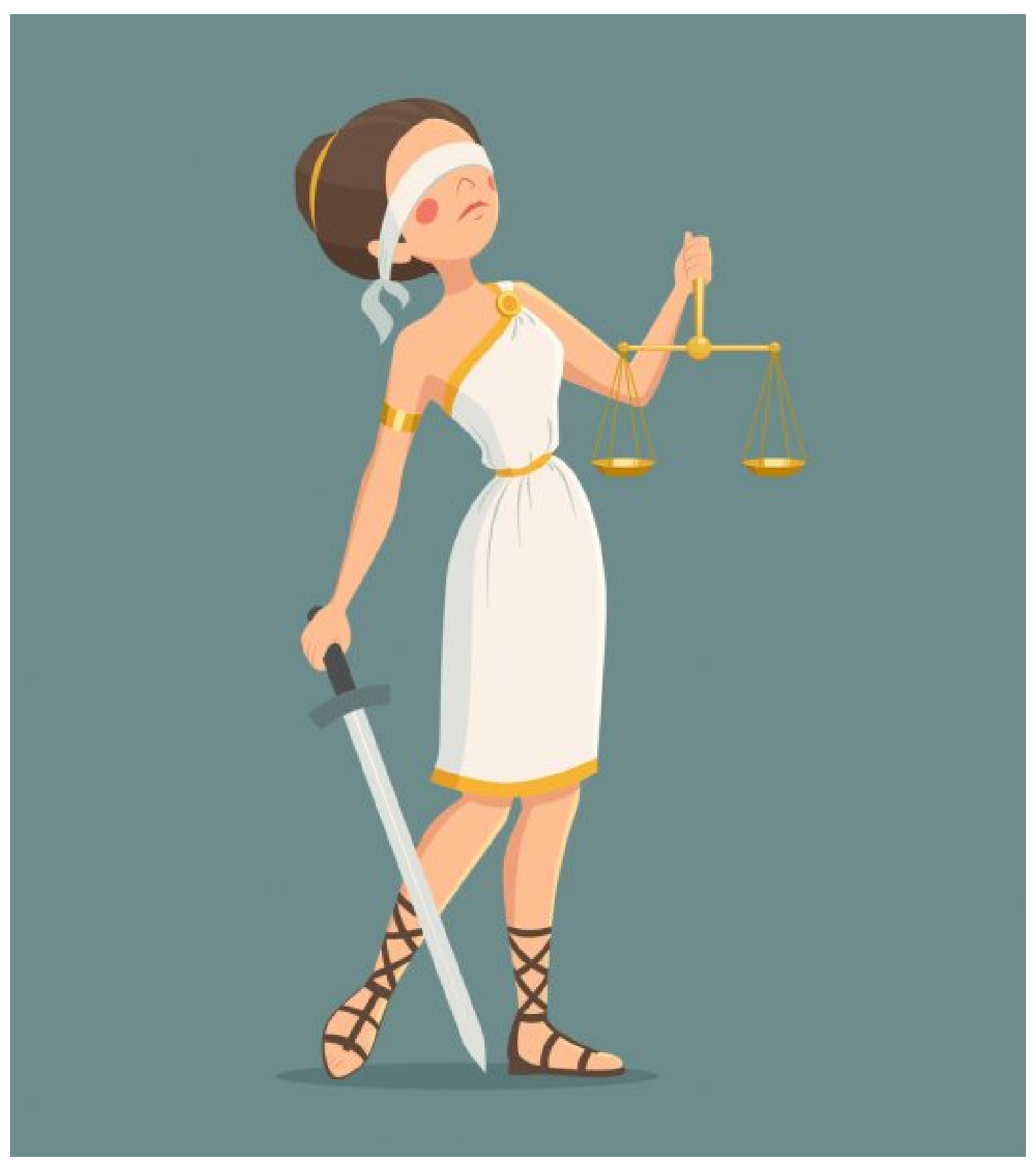
https://miro.medium.com/max/1400/1*TKti6503Vf8l83JHdGHeNA.png
https://www.terrificfitness.com.au/wp-content/uploads/2018/04/Self-sabotage.jpeg



https://miro.medium.com/max/1060/1*cc8OWxqKFXje4d_1eYrQkg.jpeg

What is the solution?

- More insights in the data.
- More balanced and fair datasets.
- More awareness during annotations.
- Unveil the black box.
- Models evaluated for bias and fairness.
- Mitigate biases in trained models.



https://st2.depositphotos.com/2885805/9312/v/
600/depositphotos_
93121394-stock-illustration-justice-lady-illustration.

My plan!

- Identify biases in large scale language models (NorBERT and NB-BERT).

- Identify the types of biases in data, and biases inherited in downstream tasks.

- Define how to mitigate such biases, and develop guidelines for Norwegian NLP.

References

Hovy, D., Prabhumoye, S. (2021). Five sources of bias in natural language processing. Language and Linguistics Compass.

Touileb, S., Øvrelid, L., Velldal E. Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. GeBNLP2020.

Touileb, S., Øvrelid, L., Velldal E. Using Gender- and Polarity-Informed Models to Investigate Bias. GeBNLP2021.

PARTNERS

HOST

UNIVERSITY OF BERGEN

FUNDER

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).