“Yeah well, that’s just your opinion man”

The fallouts of abusive language detection

27 10 2020

Zeerak Talat
University of Sheffield | Digital Democracies Institute
Twitter: @zeerakw

Clean up the Internet

Clean up the internet is an independent, UK-based organisation concerned about the degradation in online discourse and its implications for democracy. We campaign for evidence-based action to increase civility and respect online, and to reduce online bullying, trolling, intimidation, and misinformation.

Unilever warns social media to clean up “toxic” content

Toxic Content on Digital Platfor
the rise. How Can Brands Rebu
Consumer Trust?

AMNESTY INTERNATIONAL

WHO WE ARE

TOXIC TWITTE

PLACE FOR WO

Two ways social networks could control toxic content

: The Internet’s Great Spring-

THE INTERNET OF GARBAGE

THE
NERS

Abuse, racism and hate speech - help us clean up the comments online

03

Computational Modelling

Goals of abuse detection systems

Imaginaries

- Protecting marginalized communities and people from abuse
- Encode ideological positions into labelled data
- Obtain an understanding of the subjective nature of abuse.

Realities

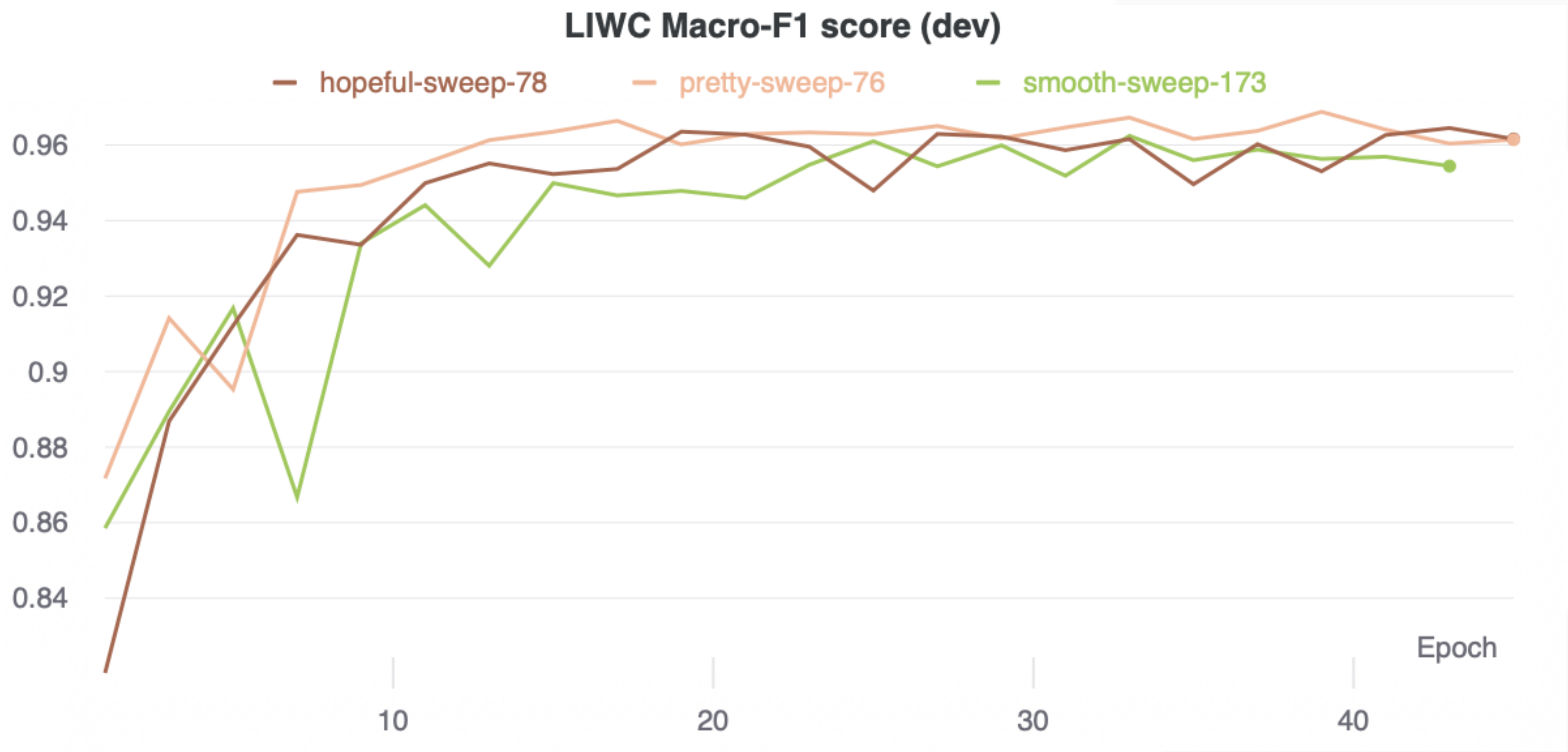
- Minority language (e.g. sociolects and dialects) use is penalized
- Use of crowd-sourcing labels as a source of ground truth
- Hardwiring (simple) correlations into the model

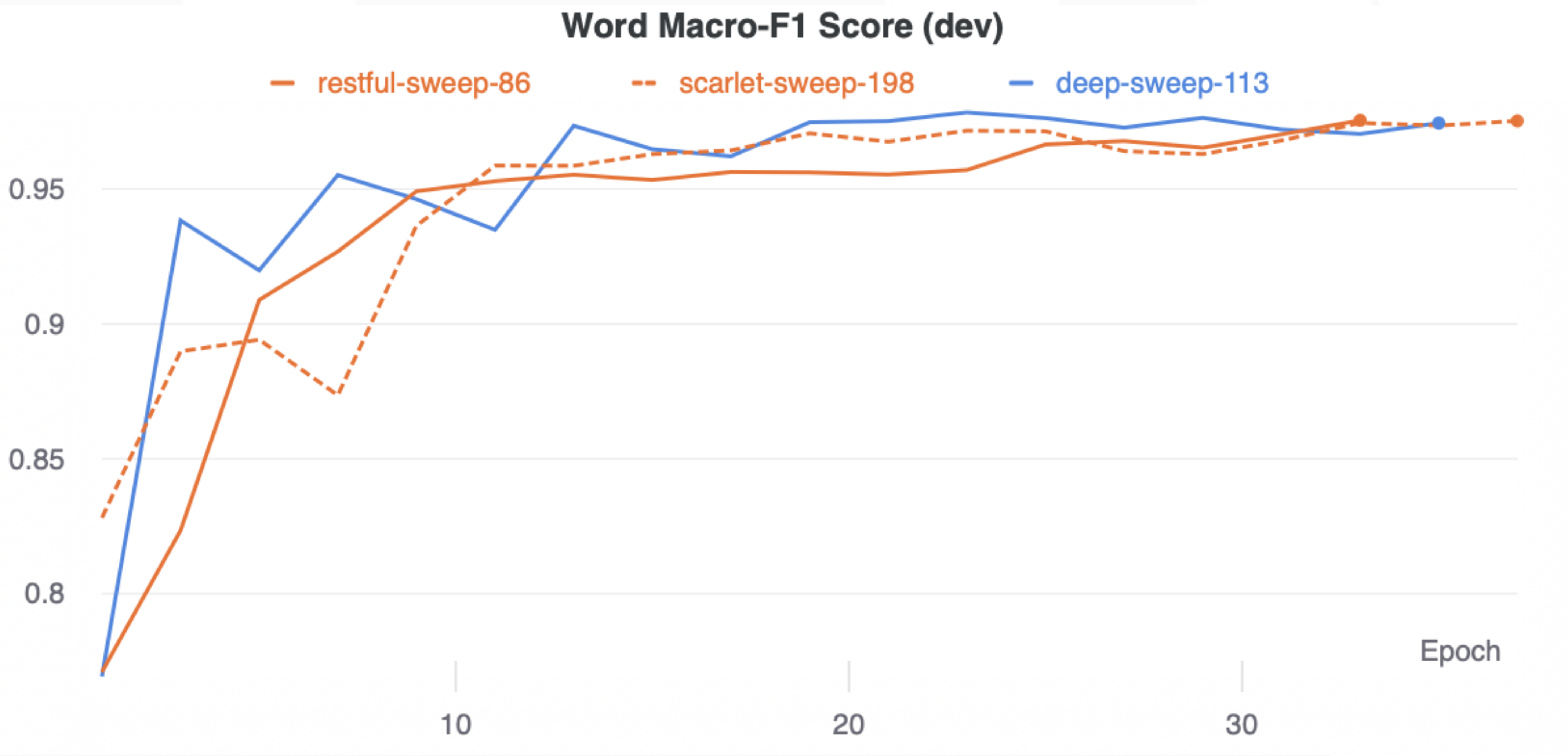
The increasing complexity of abuse detection models

- From feature engineering to inductive biases
- Pre-trained language models

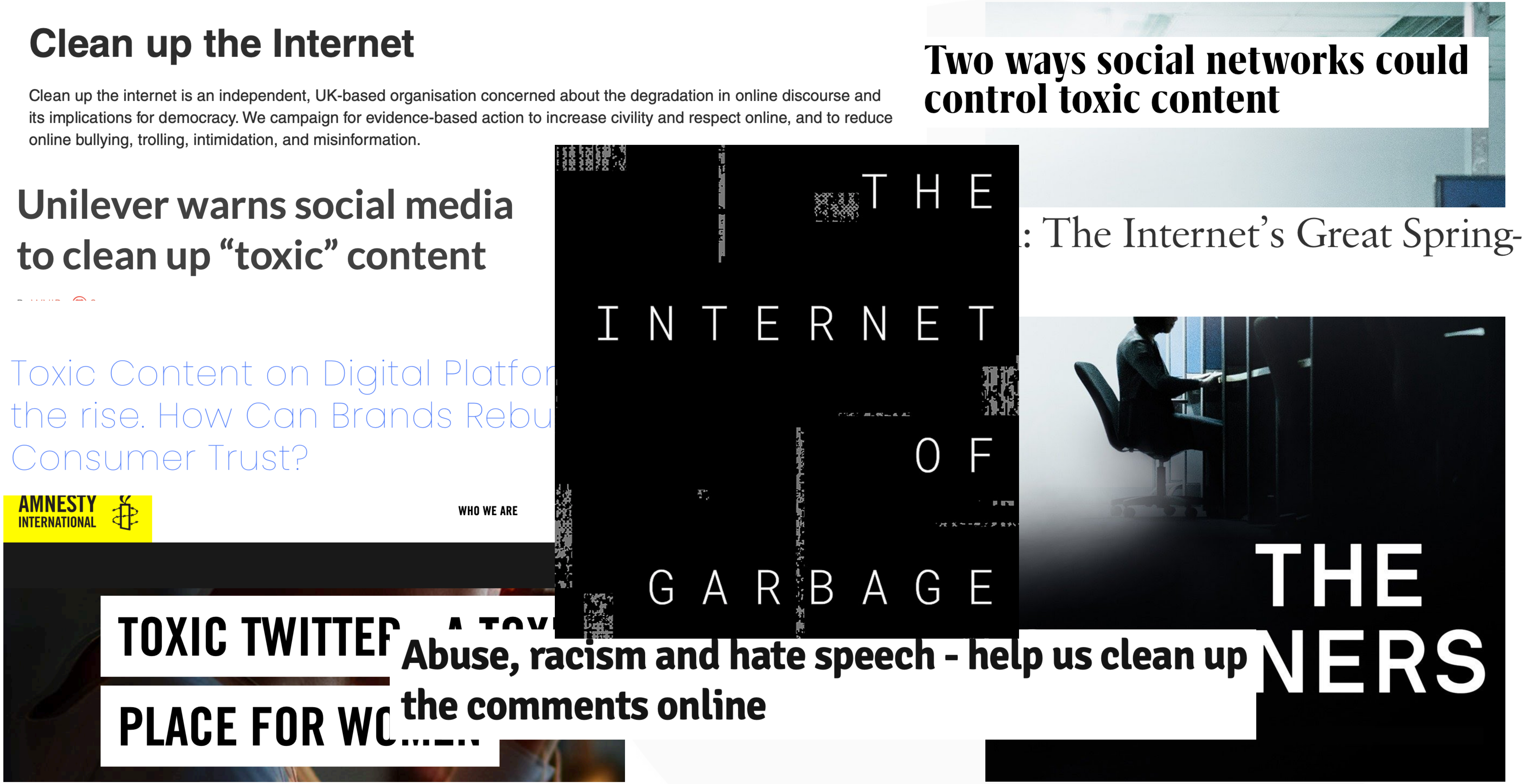
Models and inputs

Very preliminary findings from my dissertation





“Don’t you know that you’re toxic?”



Clean up the Internet

Clean up the internet is an independent, UK-based organisation concerned about the degradation in online discourse and its implications for democracy. We campaign for evidence-based action to increase civility and respect online, and to reduce online bullying, trolling, intimidation, and misinformation.

Two ways social networks could control toxic content

Unilever warns social media to clean up “toxic” content

: The Internet’s Great Spring-

Toxic Content on Digital Platfor
the rise. How Can Brands Rebu
Consumer Trust?

THE INTERNET OF GARBAGE

AMNESTY INTERNATIONAL

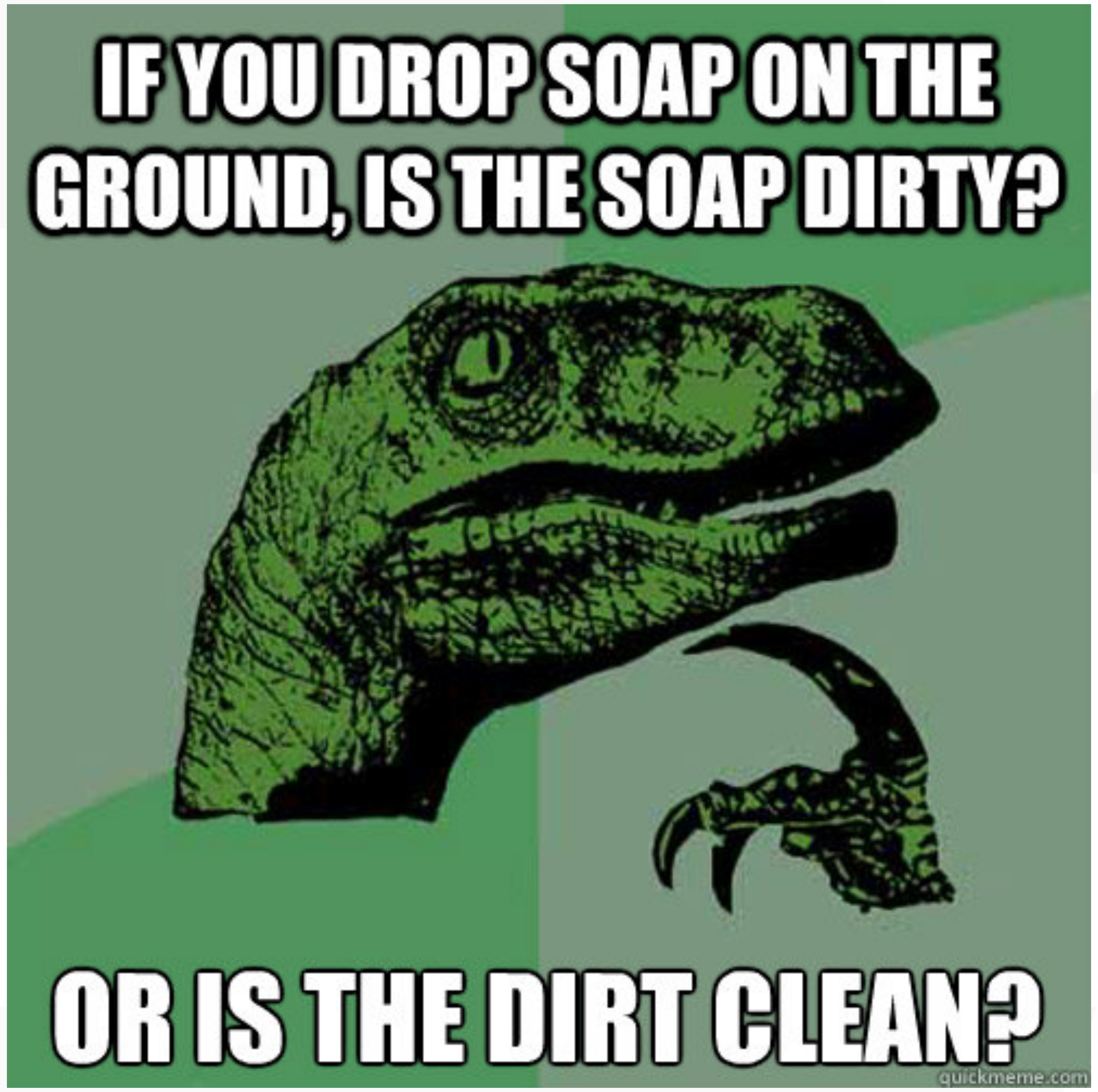
WHO WE ARE

TOXIC TWITTER - A TOX
PLACE FOR WO

THE
NERS

Abuse, racism and hate speech - help us clean up the comments online

The case of content moderation



- The "toxicity" of words
- Definitional power of "toxic"
- Independent third parties

Content moderation technology as vehicles of oppression

“No one ever accused the God of monotheism of objectivity, only of indifference”

Donna Harraway (1988),
Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.

Embodiment in machine learning pipelines

Designer

- Disassociated from data, model, and outcomes
- “Lack of diversity’ cited as reason for biased models

Data

- Collected from embodied humans, adjudicated on and disentangled from those contexts
- Subjectivity always finds its way
- Annotation == truth (to the model)
- Encode dominant discourses on objects

Future directions of content moderation technologies