Informatics DATA SCIENCE

Automated toxic comment detection

Pia Pachinger

[1]

Toxic comments on the internet

[4]

Toxic speech triggers hateful commenting behaviour and withdrawal from a debate

(Ziegele et al., 2018)

Automated content moderation

Big volume of online political talk → organisations opt for automated content moderation systems (Gillespie, 2020)

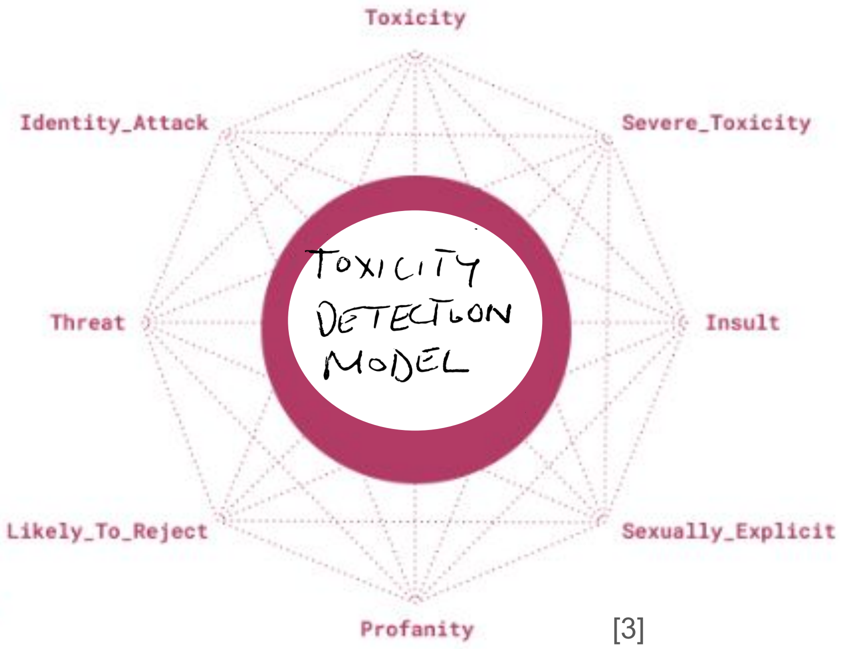
Toxic comment detection

[3]

Toxic comment detection

INPUT: TEXT

"Shut up. You're an idiot!"

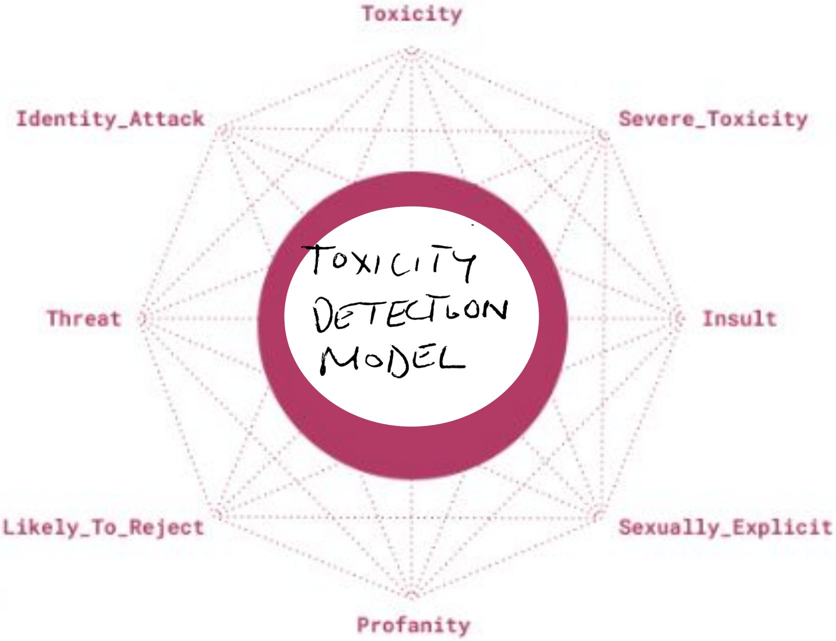


[3]

Toxic comment detection

INPUT: TEXT

"Shut up. You're an idiot!"

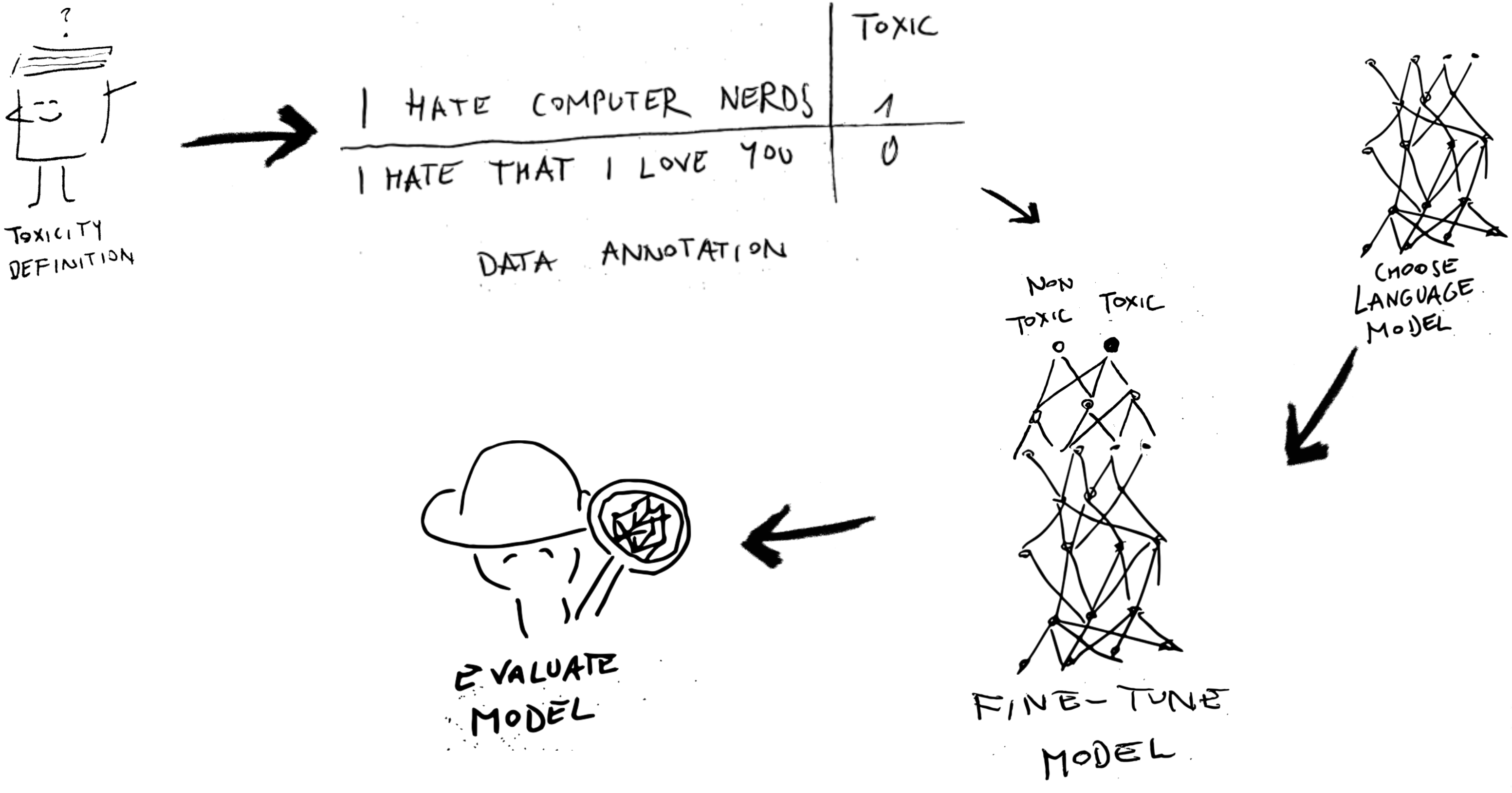


OUTPUT: SCORE

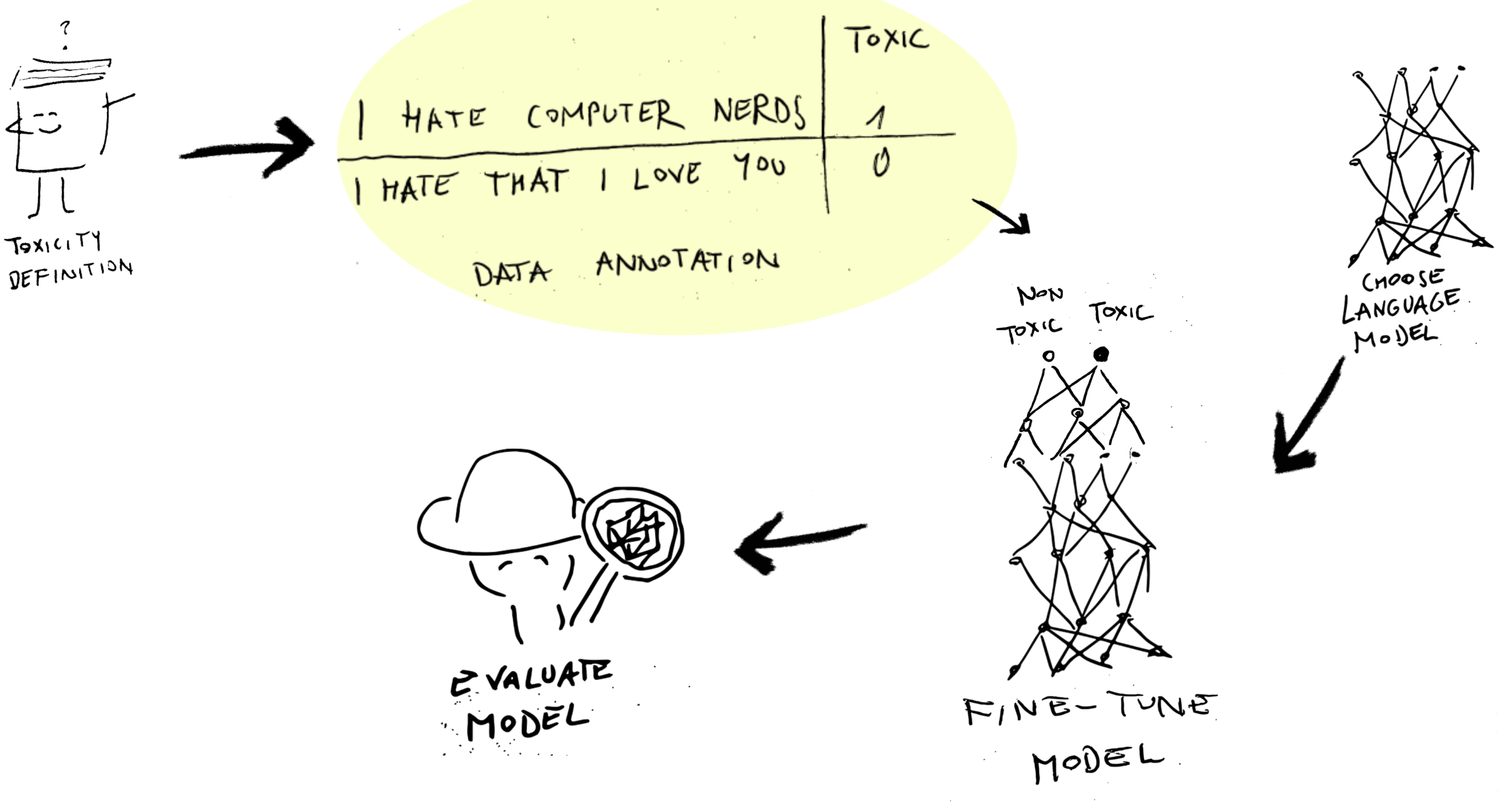
Toxicity	0.99
Severe_Toxicity	0.75
Insult	1.0
Sexually_Explicit	0.04
Profanity	0.93
Likely_To_Reject	0.99
Threat	0.15
Identity_Attack	0.03

[3]

Development of toxic comment detection model



Development of toxic comment detection model



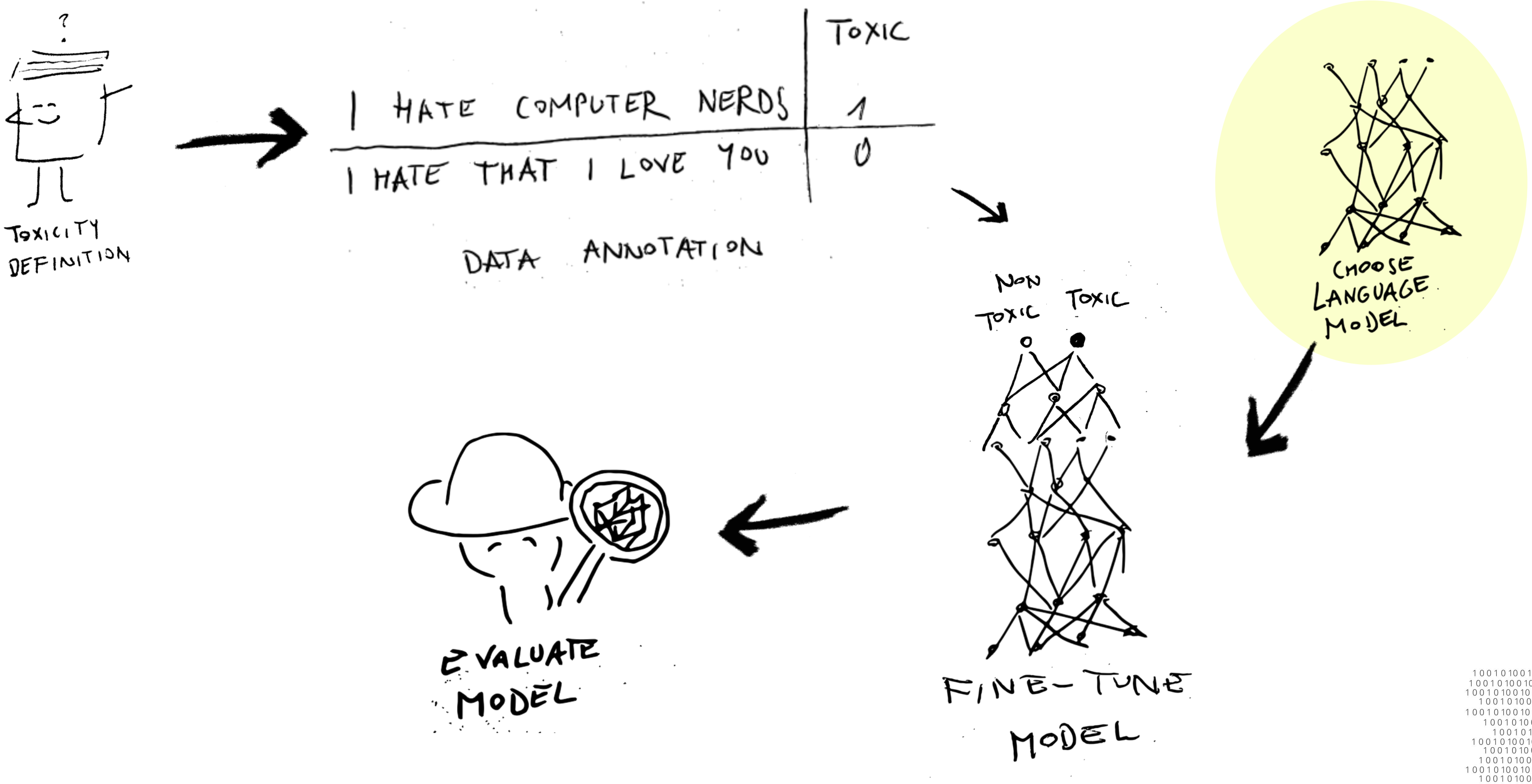
Data annotation

	comment_text	lunch_talk	love_talk	hate_talk
0	I hate that kind of food, let's not have it fo...	1	0	1
1	I hate that you love that kind of food, okay, ...	1	0	1
2	I kind of hate that I love you. Let's get lunch.	1	1	1
3	That food hates me, but I love it. I'm getting...	1	1	1

Annotation scheme: Defines how data has to be annotated by humans

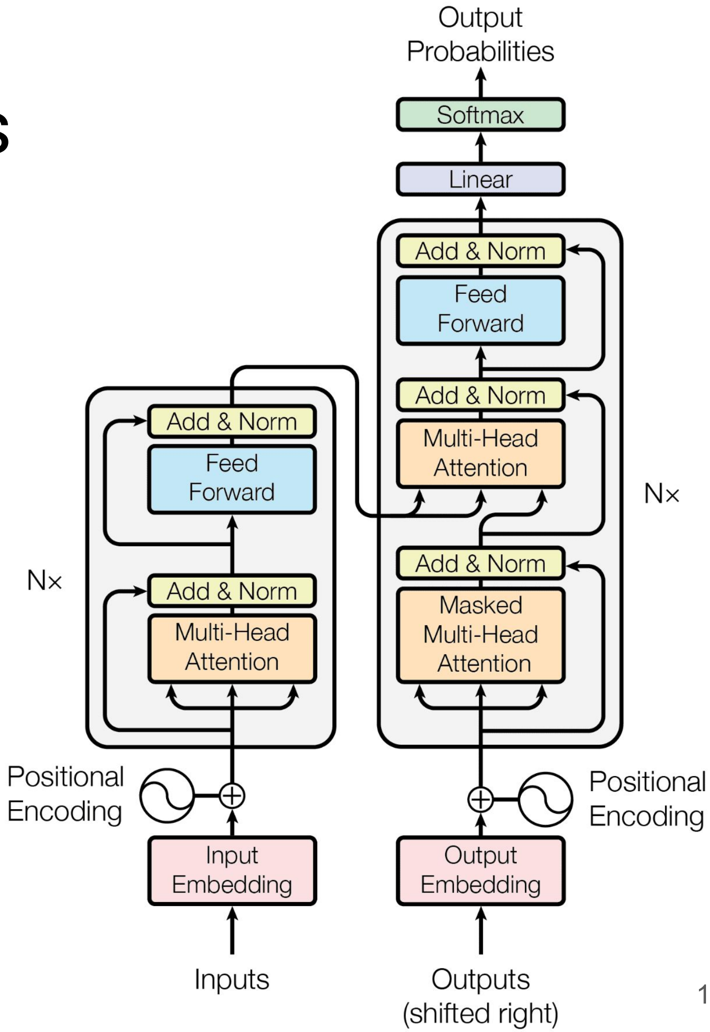
Category	Vector	Definition	Example
2. Derogation	2.1 Descriptive attacks	Characterising or describing women in a derogatory manner. This could include, but not limited to: negative generalisations about women’s abilities, appearance, sexual behaviour, intellect, character, or morals.	Women's football is so shit, they're so slow and clumsy
	2.2 Aggressive and emotive attacks	Expressing strong negative sentiment against women, such as dislike, disgust, or hatred. This can be through direct description of the speaker’s subjective emotions, baseless accusations, or the use of gendered slurs, gender-based profanities and gender-based insults.	I hate women
	2.3 Dehumanising attacks and overt sexual objectification	Derogating women by comparing them to non-human entities such as vermin, disease or refuse, or overtly reducing them to sexual objects.	Women are pigs

Development of toxic comment detection model



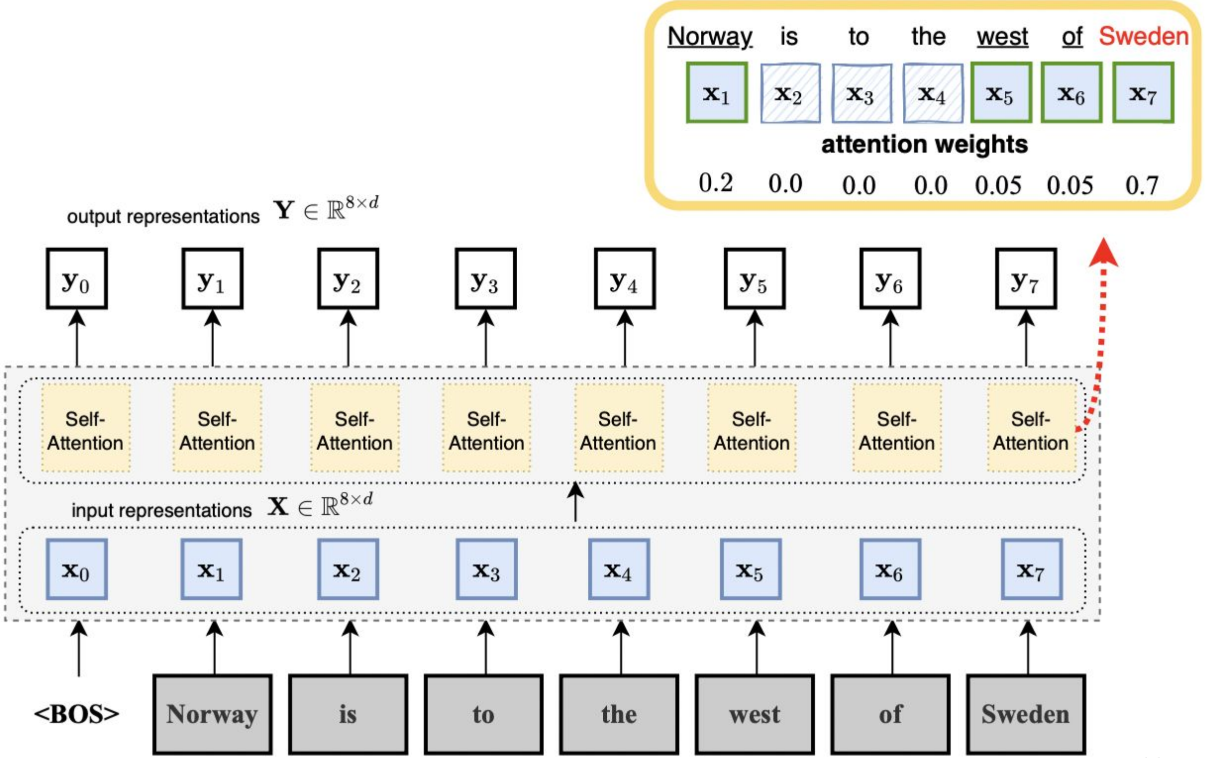
State-of-the-art language models

- Transformer based models
  - For example BERT by Google



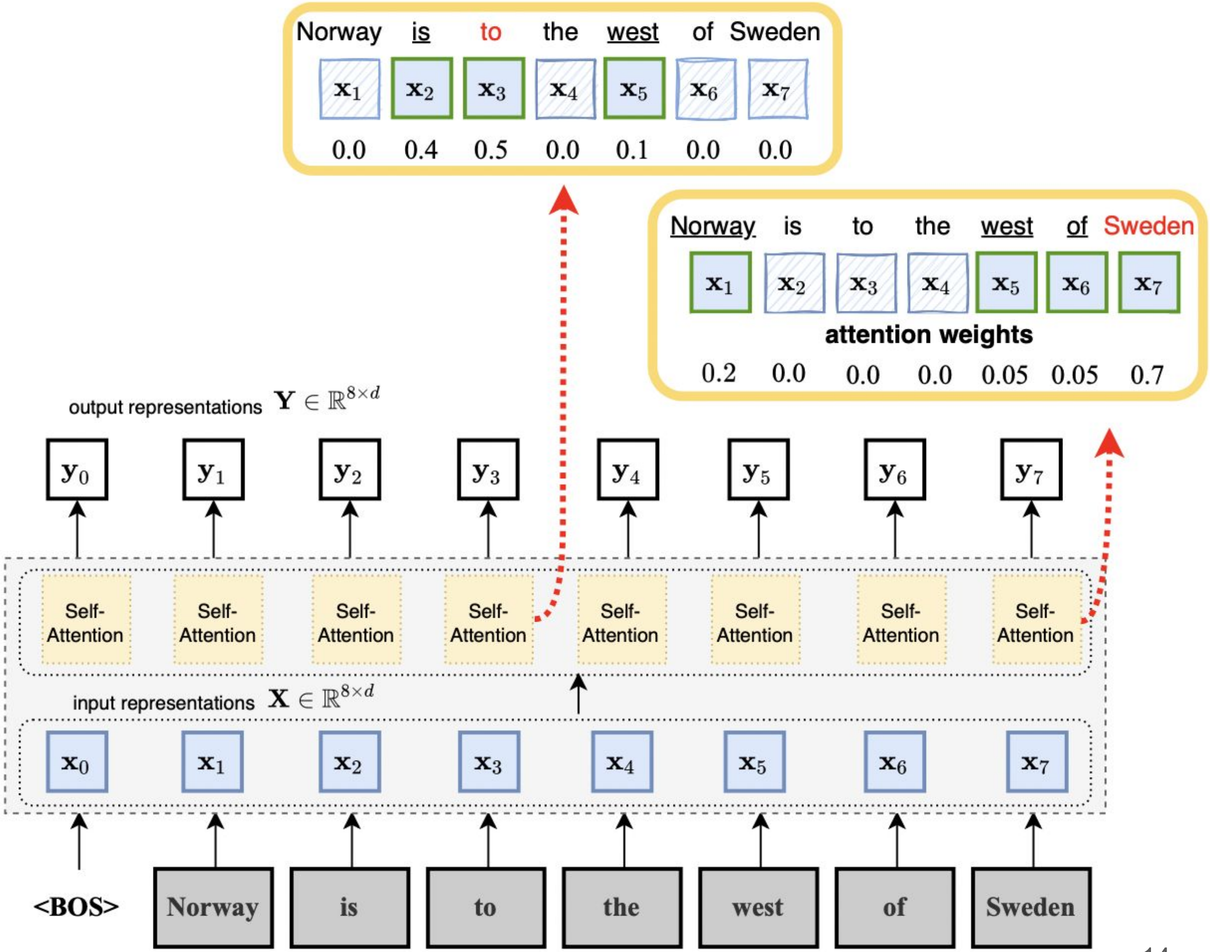
Transformers

Transformers take the <u>context</u> of a word appearing in a text into account

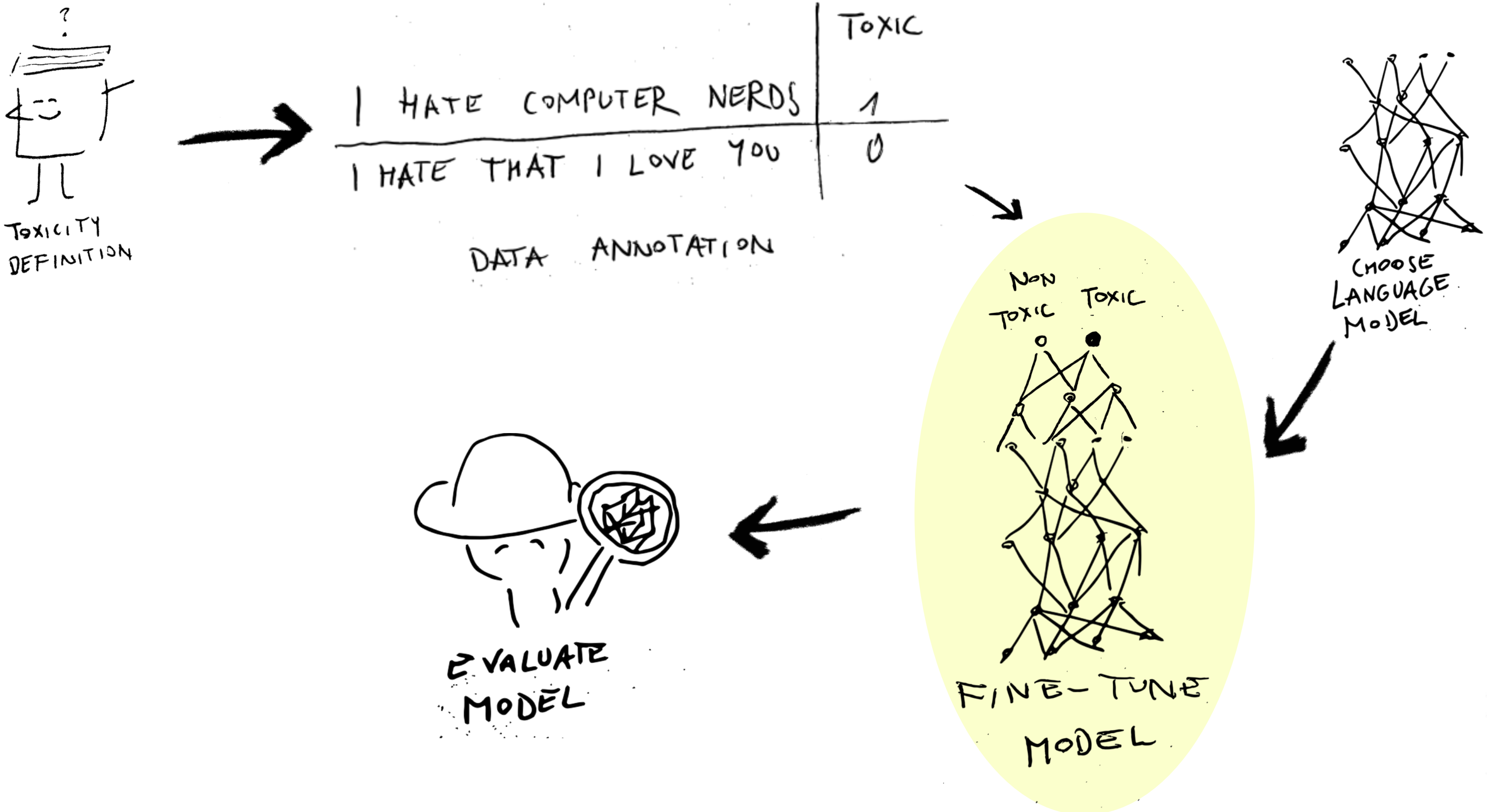


Transformers

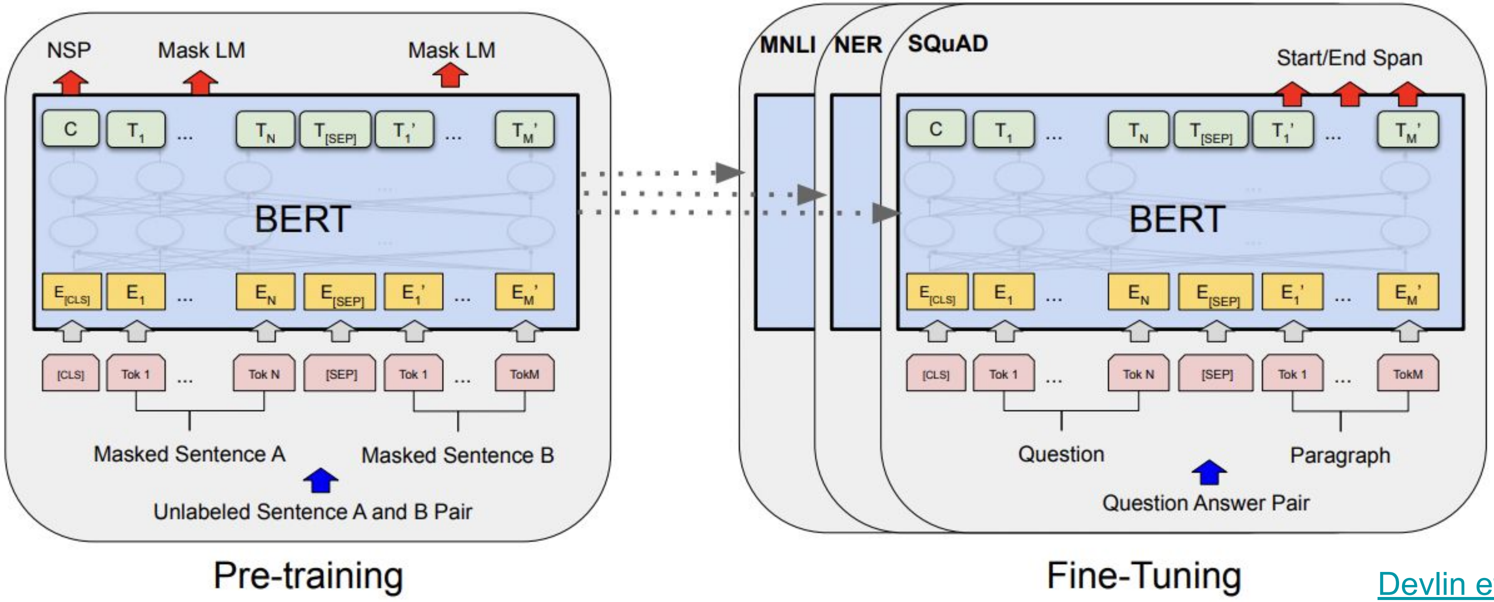
Transformers take the context of a word appearing in a text into account



Development of toxic comment detection model

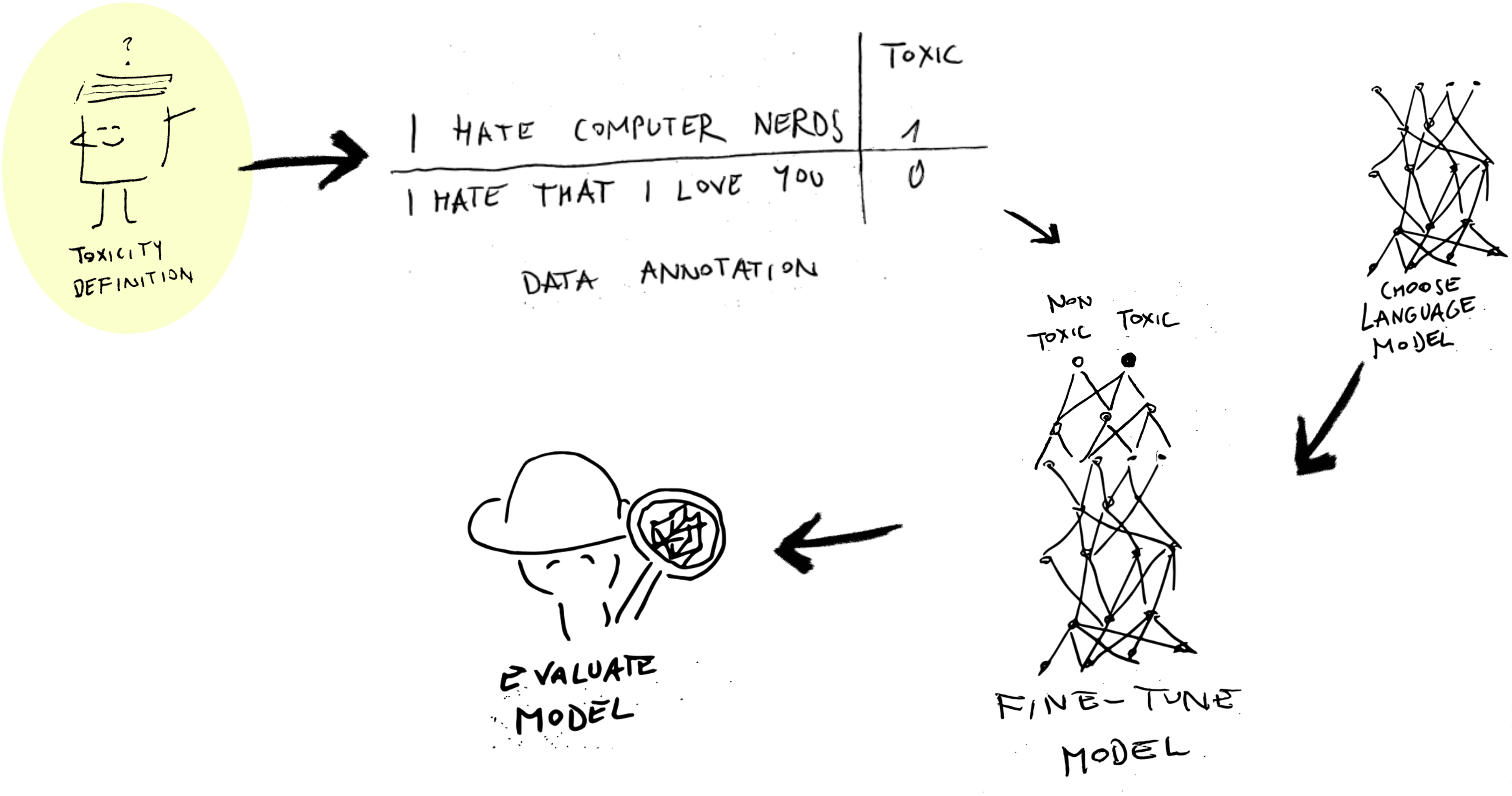


Fine-tuning general purpose models



Devlin et al., 2018

Development of toxic comment detection model

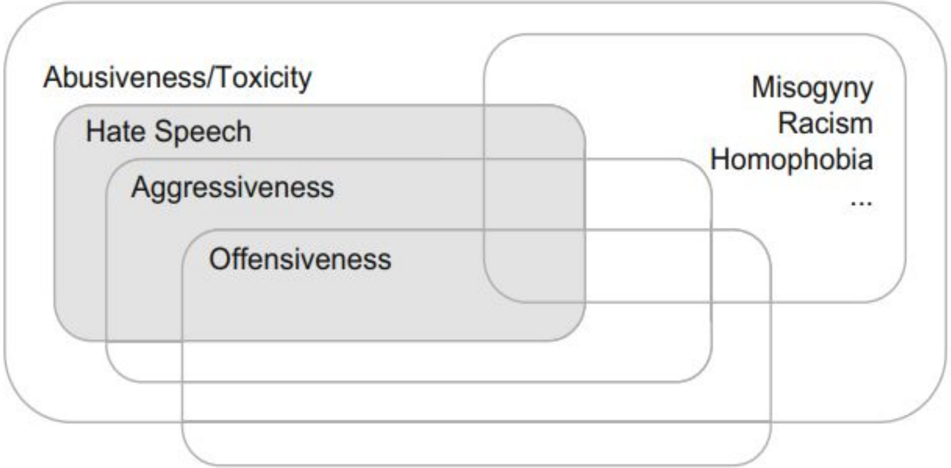


Challenges in toxic comment detection

Who defines what toxicity means?

- The perception of toxicity depends on the online community
  - For example the use of the word “nigga”
- Who defines what toxicity is?
  - Computer scientists? (Hopefully not!)

[2]



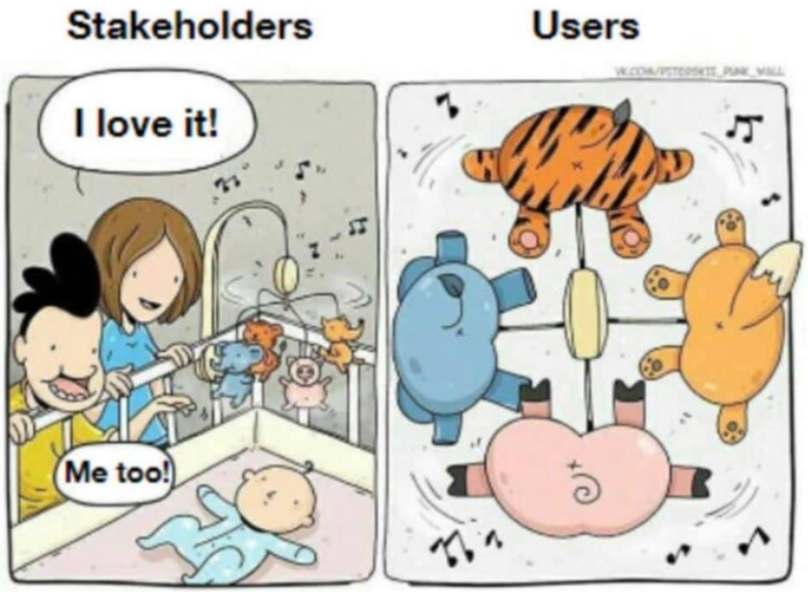
Poletto et al., 2021

TACo
Project
TRANSPARENT AUTOMATED CONTENT MODERATION

universität wien TU WIEN Informatics DATA SCIENCE derStandard.at

The TACo project

- A collaboration of the University of Vienna communication scientists and the TU Wien data scientists
- Investigates toxic language in social media from a user perspective



The TACo project - Contributions

- Investigation of the perception of toxic comments from a user perspective
    - Focus groups discuss about toxicity
    - Revision of results from focus groups via a survey
- Annotation of comments from the *DerStandard* newspaper forum
- Creation of toxicity detection model
- Evaluation of the toxicity detection model via a survey

References

[1] https://viterbischool.usc.edu/news/2020/07/context-reduces-racial-bias-in-hate-speech-detection-algorithms/

[2] https://www.governing.com/now/tension-between-online-hate-speech-and-preserving-free-speech.html

[3] Google Jigsaw

[4] https://researchoutreach.org/articles/hate-speech-regulation-social-media-intractable-contemporary-challenge/

[5] https://penpoin.com/added-value/

[6] https://www.google.com/search?q=user+perspective&sxsrf=AJOqlzXaWek0z3Wax5aagPr51J4-oPCeWQ:1674511507378&source=lnms&tbm=isch&sa=X&ved=2ahUKEwi56ca82d78AhU-q5UCHZagDhgQ_AUoAXoECAEQAw&biw=1249&bih=727&dpr=2#imgrc=vv-iNlm9lmjSHM