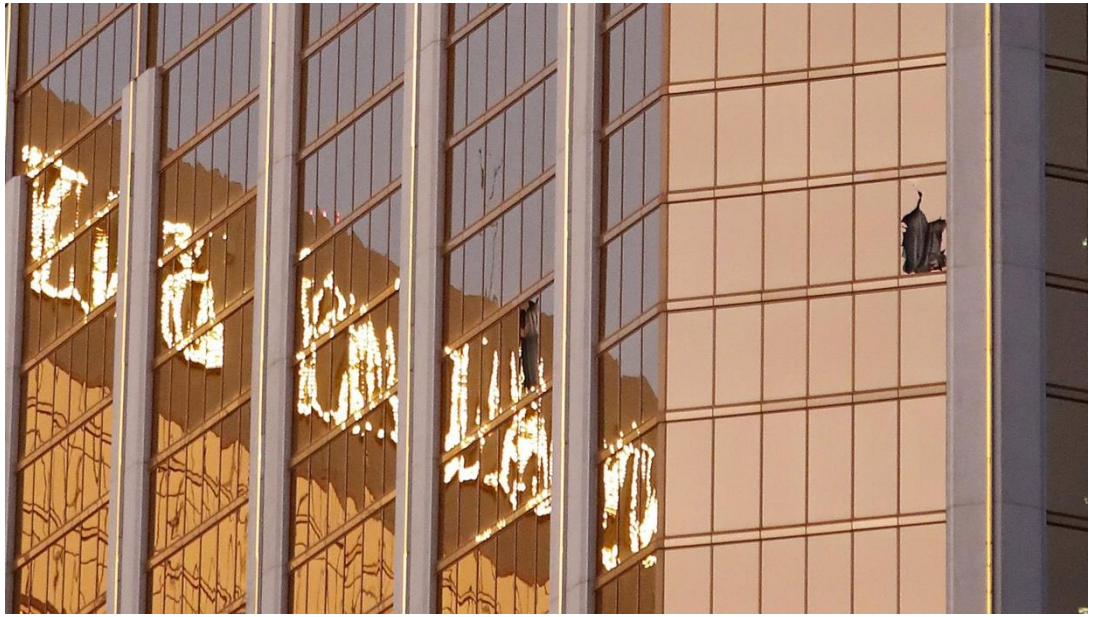
Utrecht University

Using NLP and AI to understand and prevent violence

Isabelle van der Vegt

Today's lecture

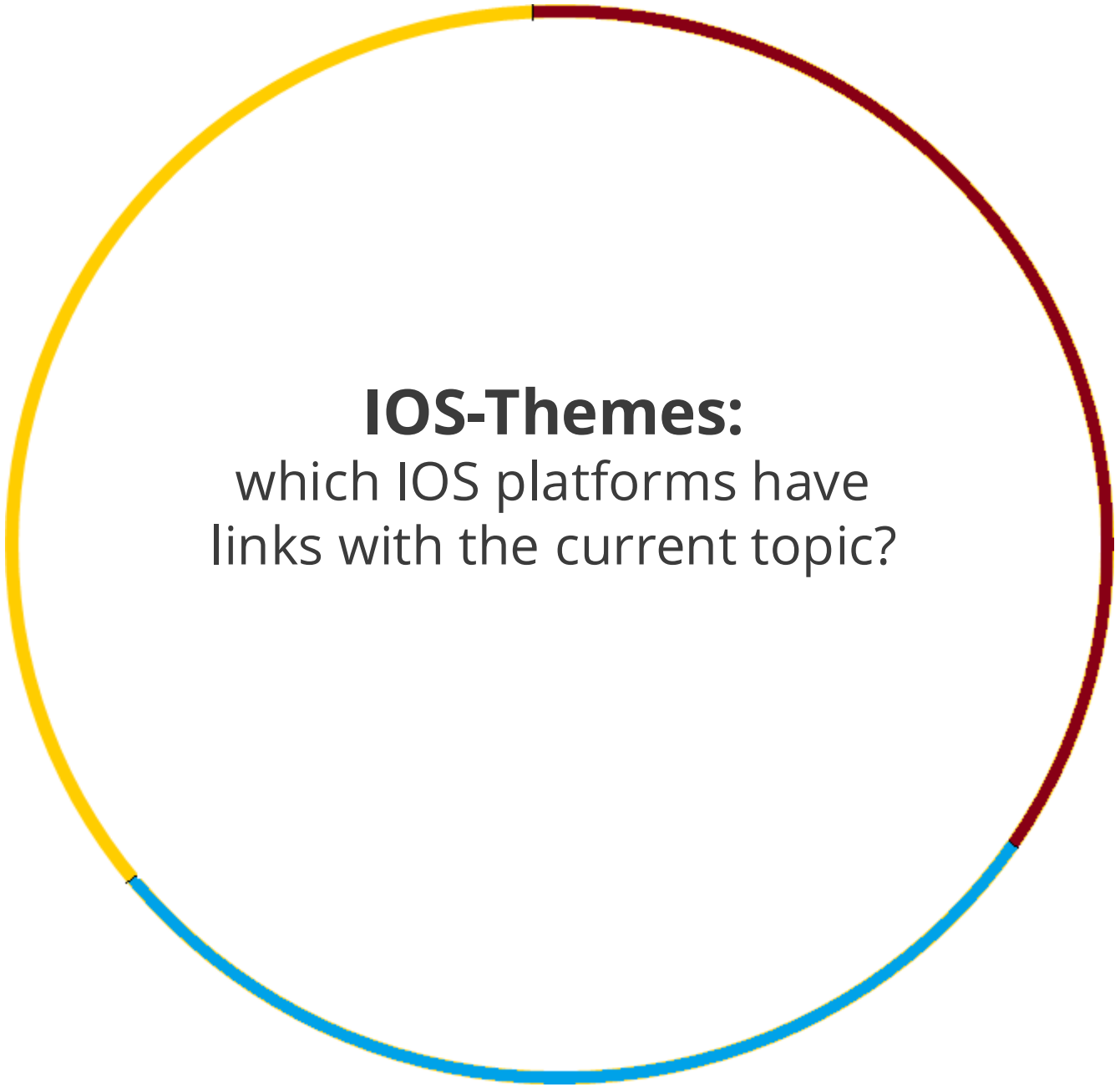


Grievance-fuelled targeted violence

Threat assessment

Natural language processing

- Large-scale text analysis
- Quantifying human language: text > numbers
- Use these ‘linguistic features’ in AI models to predict outcomes


IOS-Themes:
which IOS platforms have
links with the current topic?

Utrecht University

IOS-Themes:
15 platforms

Security in Open Societies

Bottom-up Initiatives for Societal Change

Behaviour and Institutions

Democracy and good governance

Contesting Governance

Futures of Democracy

The Transactional State as an Institution for Good

Future of Work

Open Cities

Transitions and well-being

Fair Transitions

Longtermism and Institutional Change

Markets and Corporations

Openness challenged: the university at risk?

Equality and diversity

In/Equality

Gender, Diversity and Global Justice

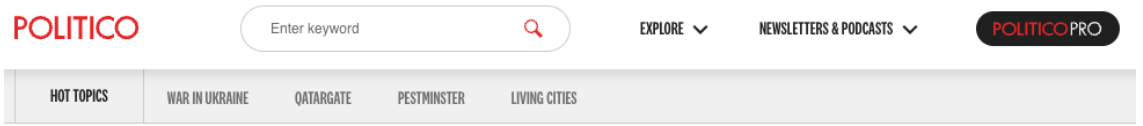
Today's objectives

- Gain insight into different NLP methods through its application in understanding and preventing violence
- Discuss the challenges and opportunities of using these methods in practice (close protection and security)
- Examine the (possible) impact of these methods on society

Understanding threats of violence

Study 1: threats to politicians

- NLP method: supervised machine learning
- What is the prevalence and nature of abuse directed at politicians online in the Netherlands?
- Is there an effect of the gender and ethnic minority status of politicians on the prevalence and nature of abuse?



Top Dutch minister steps down from party leadership over 'intimidation' and 'threats'

Research questions

What is the prevalence and nature of abuse directed at politicians online in the Netherlands?

Is there an effect of the gender and ethnic minority status of politicians on the prevalence and nature of abuse?

Data

X (Twitter) data

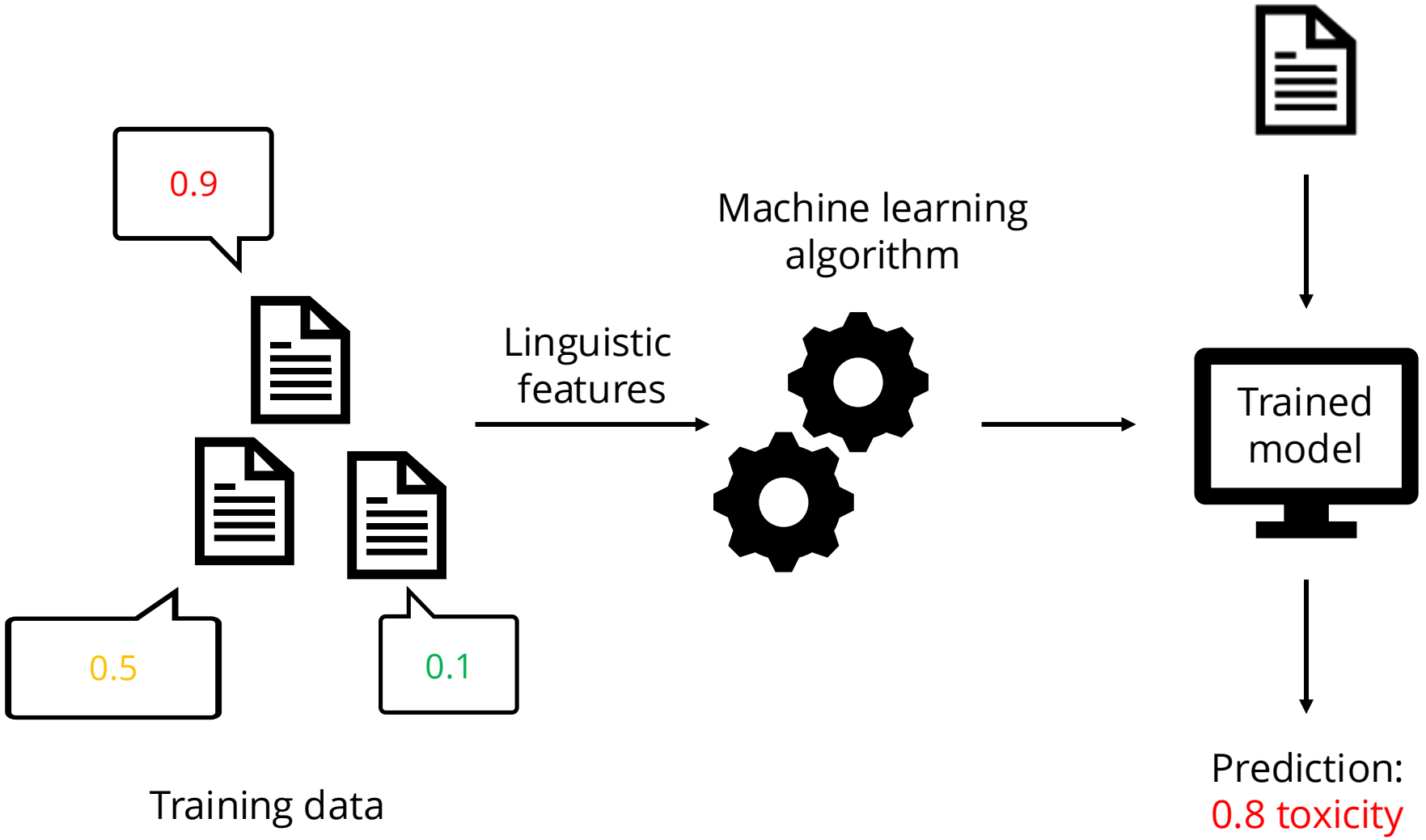
- @Mentions of all Dutch party leaders (*n* = 22) in 2022
- Total 1,909,844 tweets collected
- Twitter Academic API, using academictwitteR package

Method

Linguistic measures

- Google Perspective API: supervised machine learning models
- Score of 0-1, where 1 = 100% of people would agree

1. Toxicity
2. Severe toxicity
3. Identity attack
4. Insult
5. Profanity
6. Threat


0.9
0.5
0.1
Training data
Linguistic features
Machine learning algorithm
Trained model
Prediction:
0.8 toxicity

Dutch (original)	English	Toxicity	Severe toxicity	Identity attack	Insult	Profanity	Threat
hang jezelf aub op, vieze kanker jood. Je verneukt heel nederland, hoop dat je word doorzeefd met kogels	Please hang yourself, dirty “cancer” jew. You’re ruining the Netherlands, hope you get riddled with bullets	0.96	0.82	0.84	0.93	0.95	0.86
Smerige KutHoer,,, kogel in je kop moet je krijgen	Dirty “cunt” “cancer” whore,,, you should get a bullet in the head	0.93	0.82	0.11	0.87	0.89	0.85
Rot jij maar lekker op naar je vaderland met je klote islam.	Piss off to your homecountry with your damn islam.	0.92	0.82	0.85	0.80	0.76	0.35
Domme muts ben je. Serieus wat een achtelijk schijtwijf. Geen andere woorsen voor deze domme domme domme opmerking van je.	Stupid bimbo you are. Seriously what a retarded “crap” woman. No other words for this stupid stupid stupid remark of yours.	0.89	0.66	0.06	0.88	0.84	0.02

Method

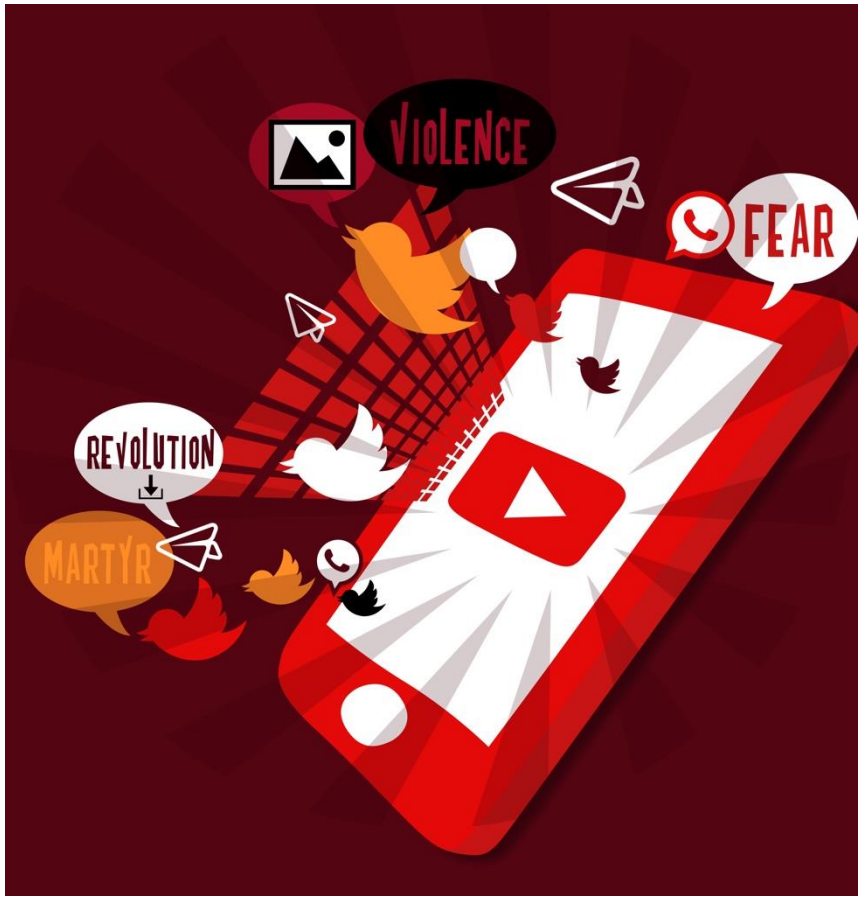
Regression model

- Gender
- Ethnic minority status

- Political position: economic and cultural stance
- Number of followers
- Number of tweets

Results

- **Male politicians score higher than female politicians** on toxicity, severe toxicity, identity attacks, insults, and profanity. **No significant differences found for threats.**
- Significant **interactions between gender and ethnic minority status** were found for for severe toxicity, identity attacks, profanity, and threats.
- Tweets directed at female politicians with ethnic minority background: **most threatening**



Utrecht University

Paper: https://rdcu.be/edKB1

Discussion

- Google Perspective (and similar AI-based tools) are used for content moderation across platforms
- The problem of using proprietary AI tools: not transparent and sometimes biased, for example..

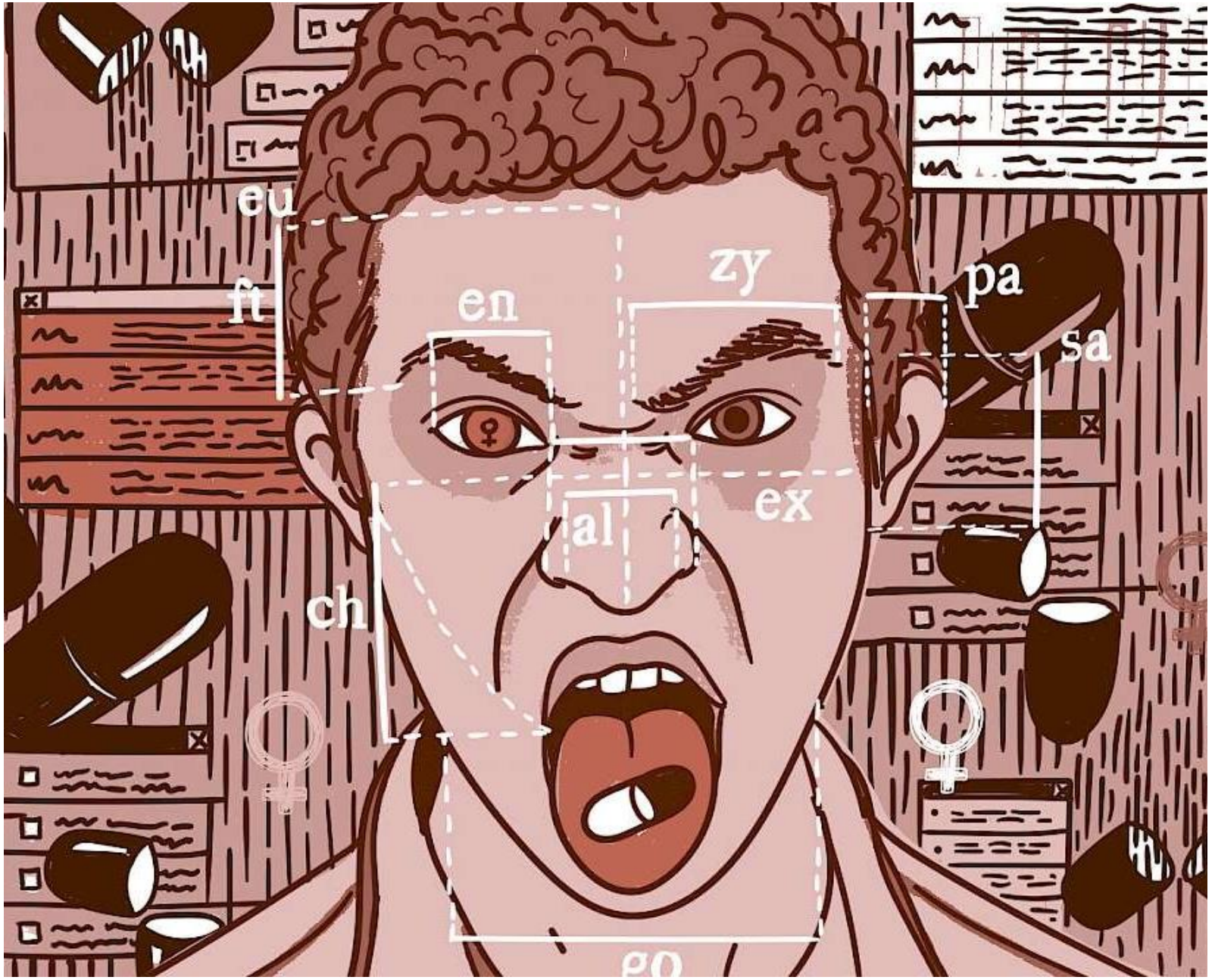
Identity attacks measure

‘Negative or hateful comments targeting someone because of their identity, including but not limited to race or ethnicity, religion, **gender,** nationality or citizenship, disability, age, or sexual orientation.’

Dutch (original)	English	Gender	Identity attack
hang jezelf aub op, vieze kanker jood. Je verneukt heel nederland, hoop dat je word doorzeefd met kogels	Please hang yourself, dirty “cancer”* jew. You’re ruining the Netherlands, hope you get riddled with bullets	M	0.84
Smerige KutHoer,,, kogel in je kop moet je krijgen	Dirty “cunt” “cancer” whore,,, you should get a bullet in the head	F	0.11
Rot jij maar lekker op naar je vaderland met je klote islam.	Piss off to your homecountry with your damn islam.	M	0.85
Domme muts ben je. Serieus wat een achtelijk schijtwijf. Geen andere woorsen voor deze domme domme domme opmerking van je.	Stupid bimbo you are. Seriously what a retarded “crap” woman. No other words for this stupid stupid stupid remark of yours.	F	0.06

Study 2: Cross-platform analysis of incel language

- NLP method: dictionaries
- Is incel subculture a violent extremist ideology?
- Has language grown more extreme over time?



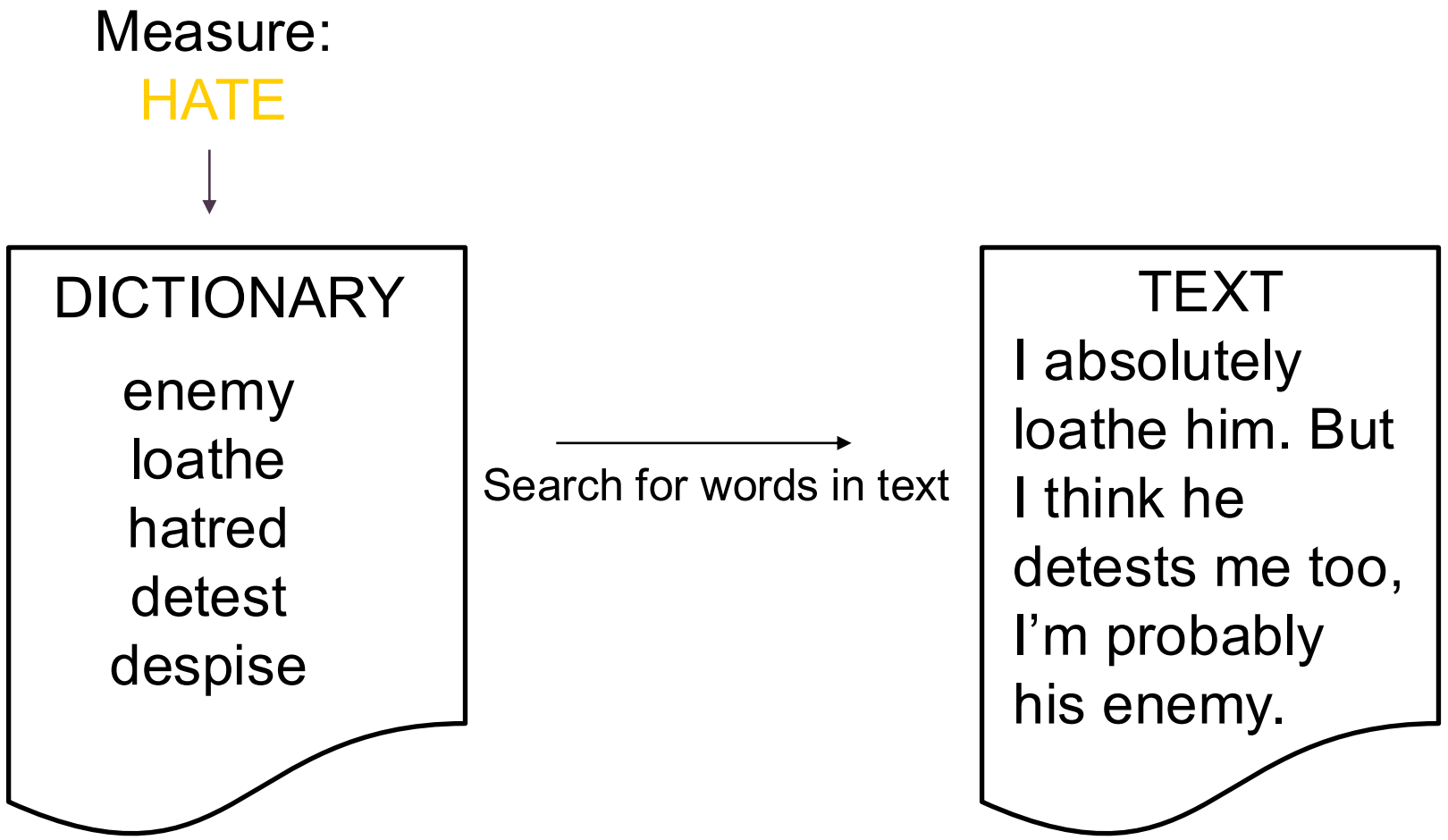
TERRORISM AND POLITICAL VIOLENCE
2024, VOL. 36, NO. 3, 382–405
https://doi.org/10.1080/09546553.2022.2161373

Routledge
Taylor & Francis Group

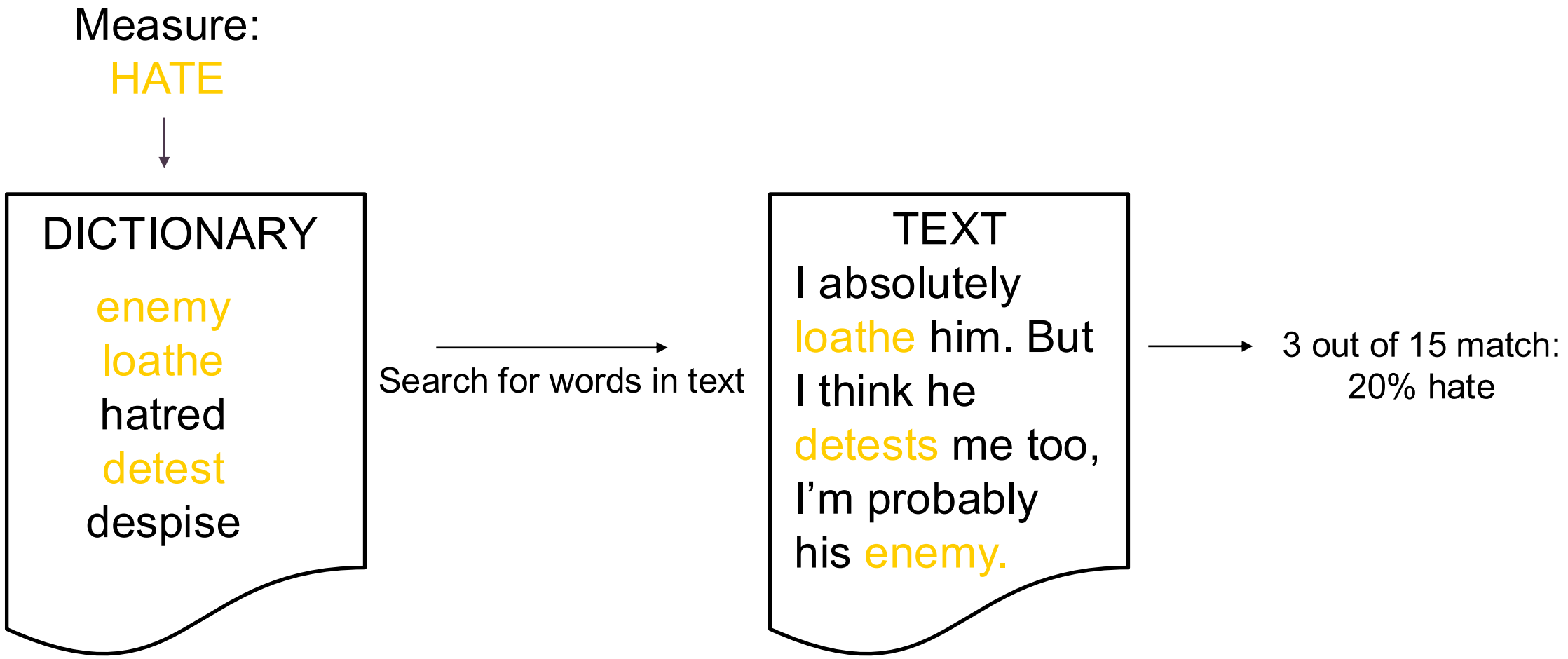
A Diachronic Cross-Platforms Analysis of Violent Extremist Language in the Incel Online Ecosystem

Stephane Baele[a], Lewys Brace[b], and Debbie Ging[c]

Dictionary-based NLP



Dictionary-based NLP



Incel violent extremist dictionary

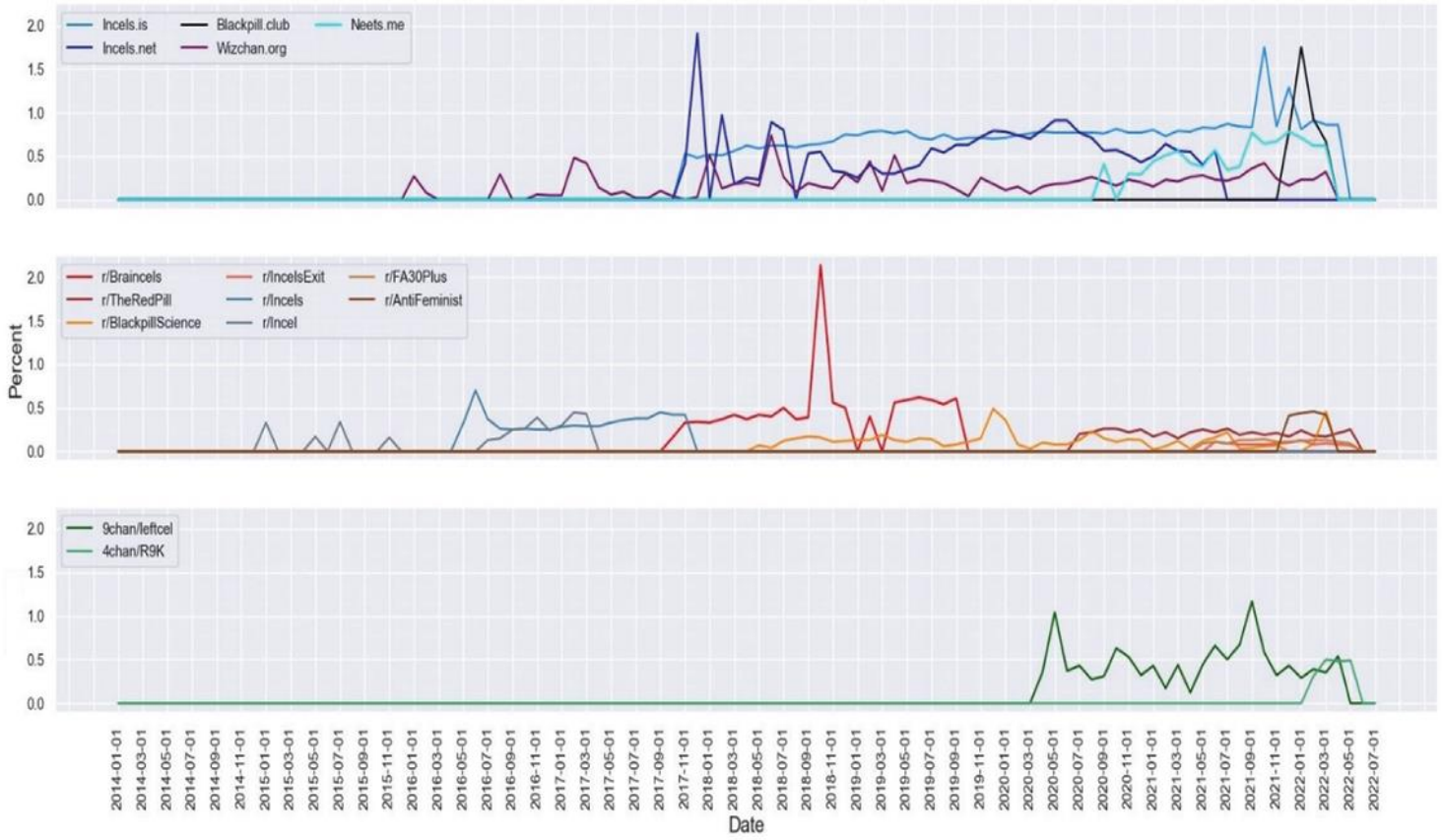
- Verbs unambiguously expressing acts of violence (e.g., “stab,” “kill,” “rape”)
- Nouns that label weapons (e.g., “gun,” “knife,” “acid”),
- Nouns that dehumanize the outgroups (e.g., “femoid”/“foid,” “roasties”; “curry”)

- Total 172 words judged by experts on incelosphere

Data

- 33 platforms: subreddits, forums, Chan image boards, blogs, Telegram channels
- Custom scrapers to collect total 11,717,516 posts

Results


Incels.is
Incels.net
Blackpill.club
Wizchan.org
Neets.me
r/Braincels
r/TheRedPill
r/BlackpillScience
r/IncelsExit
r/Incels
r/Incel
r/FA30Plus
r/AntiFeminist
9chan/leftcel
4chan/R9K
Percent
Date

Discussion

Limitations of dictionary-based methods

- Labor intensive to develop and maintain
- Not aware of context, irony, sarcasm

Benefits: transparent and interpretable

New developments: LLMs

Preventing threats of violence

Threat assessment

Threat assessment

- Estimate the risk of violence, or other outcomes like seriousness and likelihood
- Teams of e.g., police, mental health, and investigative psychologists
- Structured professional judgement tools

CTAP-25

1. Threats
2. Declaration of intention
3. Evidence of displacement
4. Extremes of anger
5. Escalation in anger or increasing preoccupation
6. Highly personal quest for justice
7. Demands to change behaviour
8. Demand for money/apology
9. Prolific correspondence
10. Awareness of personal details
11. History of intrusive behaviours
12. End-of-tether language
13. Suicidal ideation
14. Sexually aggressive language or fantasies
15. Interest in attackers or violent extremism
16. References to weapons
17. Known history of violence
18. Homicidal ideation
19. Delusion of loving relationship
20. Delusions of jealousy
21. Belief in shared past or destiny
22. Belief people are imposters or posessed
23. Threat to personal integrity
24. Belief in own divinity or on a divine mission
25. Gut reaction

Level of Concern: Low / Medium / High

Utrecht University

The implications of AI for Close Protection and Surveillance

- What are the most promising applications of AI?
- What are the most pressing challenges?

- Focus on human analyses and threat assessments

Applications of AI

- Basic administrative tasks
- Open Source Intelligence (OSINT) analysis

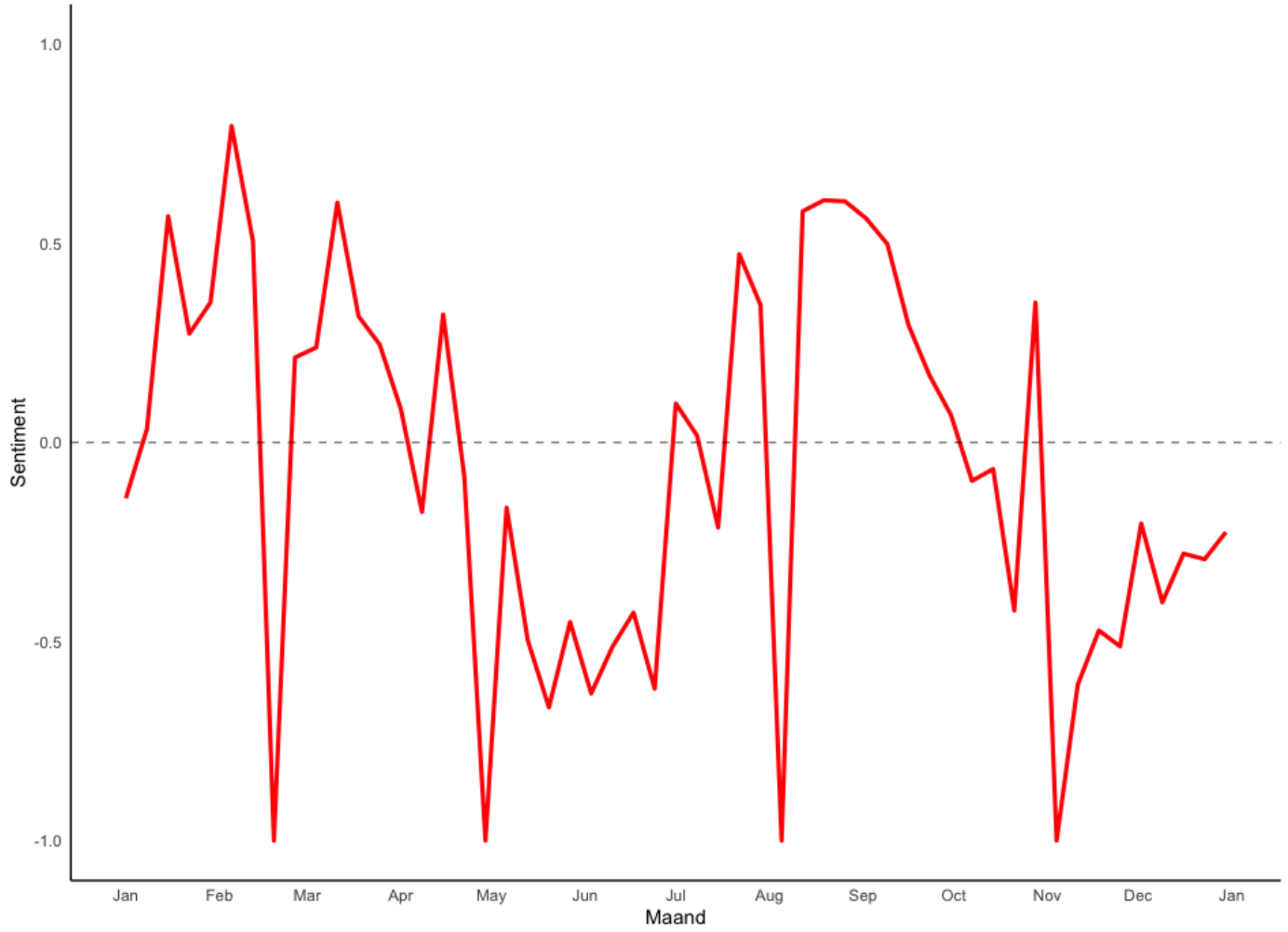
“What is the threat towards a person? What is the source? What can we expect? There is so much information that you can collect. Humans are not able to do all of that anymore. They cannot read all that information and deduce meaningful insights from all of that.”

Applications of AI: OSINT

Dashboards

- Collect, summarise, visualize and/or analyze information from social media
- Prioritize information

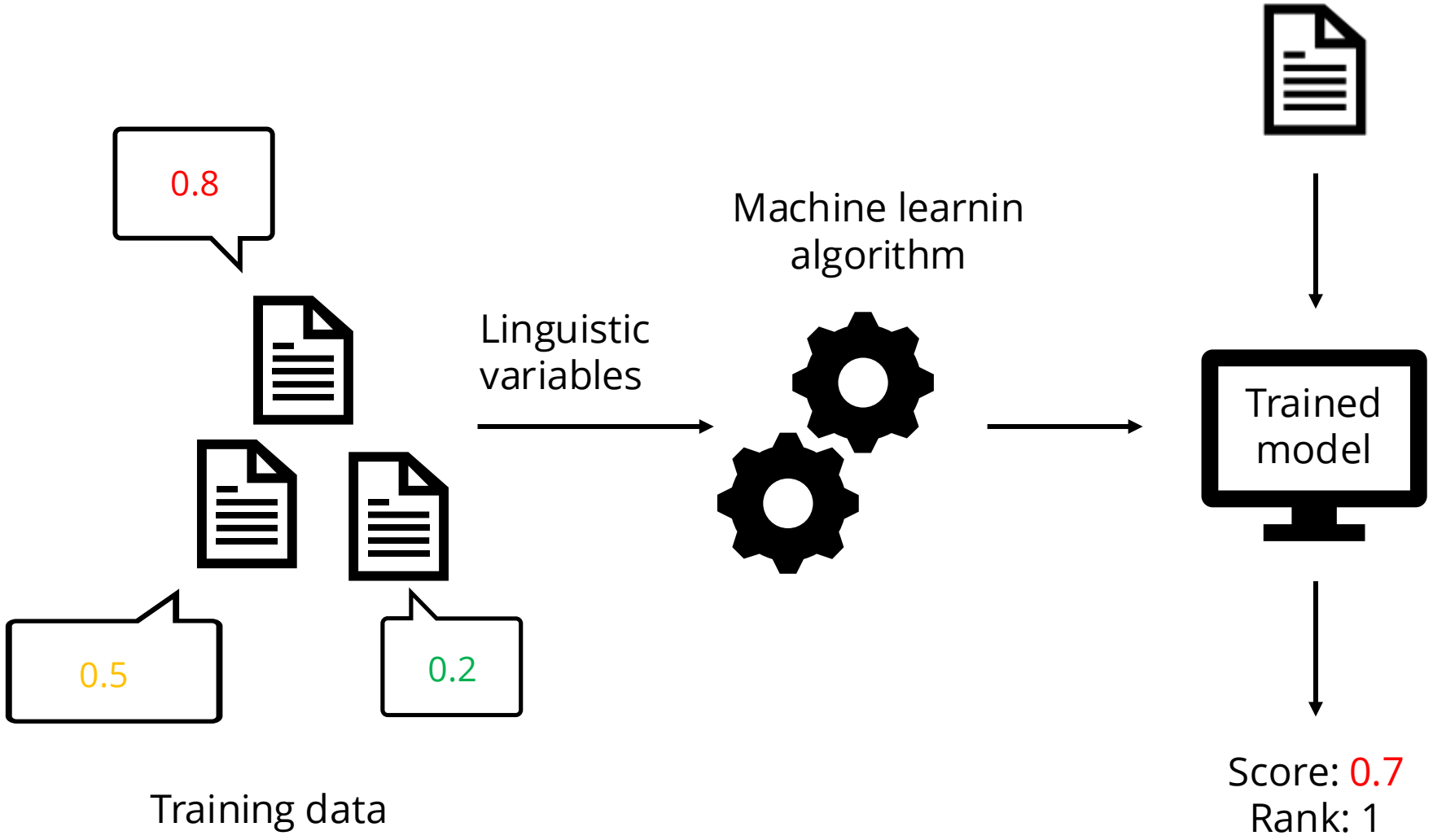
Visualize: What is the general attitude towards a public figure online?



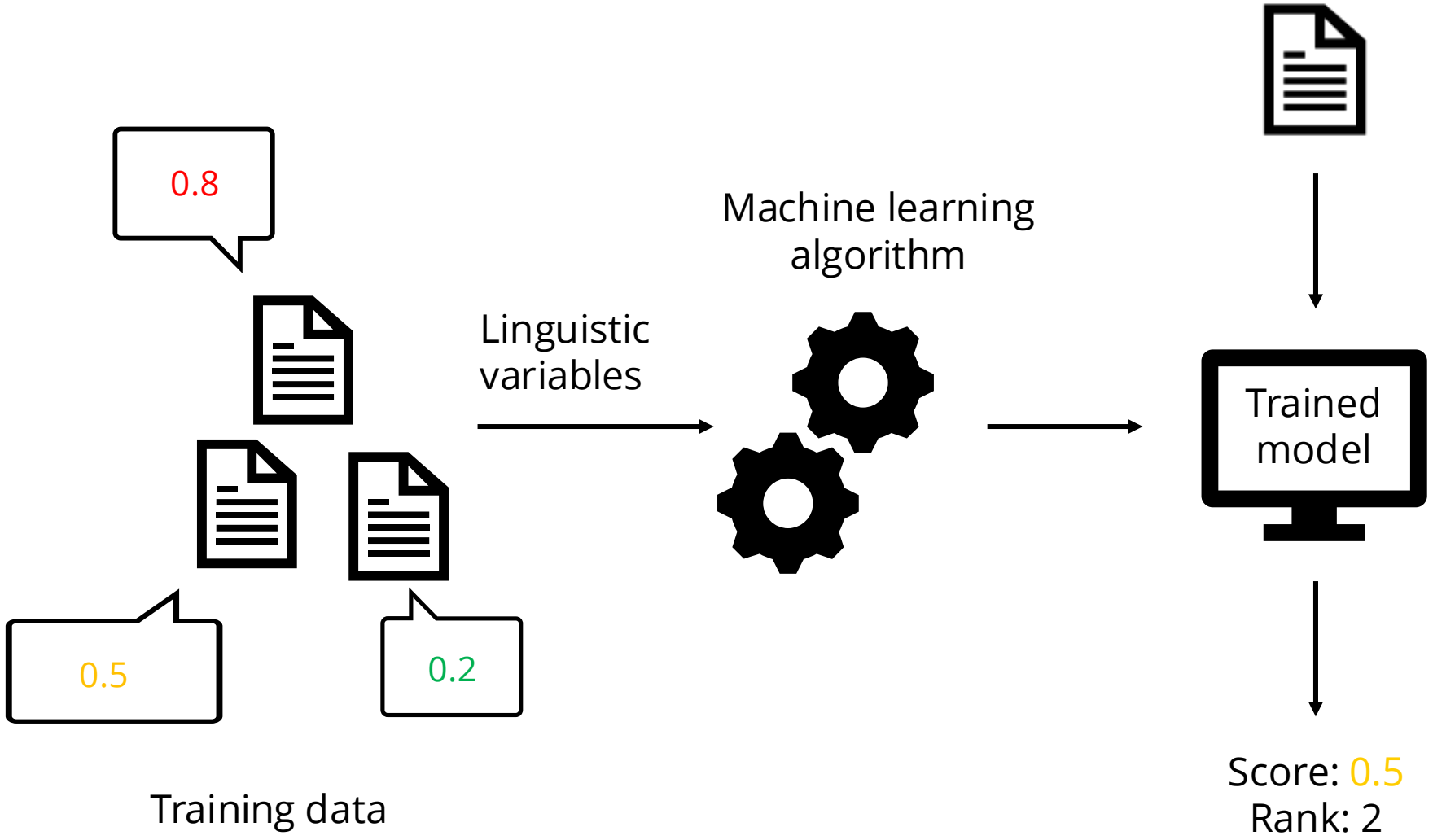
Sentiment analysis

Utrecht University

Prioritize: which messages call for violent action?



Prioritize: which messages call for violent action?



Summarize

There is an increasing number of posts with #wefpuppet (+2,230 posts). The hashtag #wefpuppet most likely refers to conspiracy theories surrounding the World Economic Forum (WEF). In this context, it is mentioned 253 times that the politician X is a “puppet” of the WEF.

Applications of AI: threat assessment

Applications of AI: threat assessment

Applications of AI: threat assessment

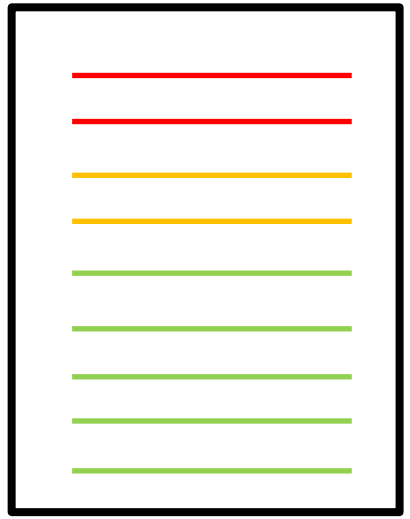
Support

- Measure existing risk factors (from SPJ) using AI

Applications of AI: threat assessment

Human-in-the-loop

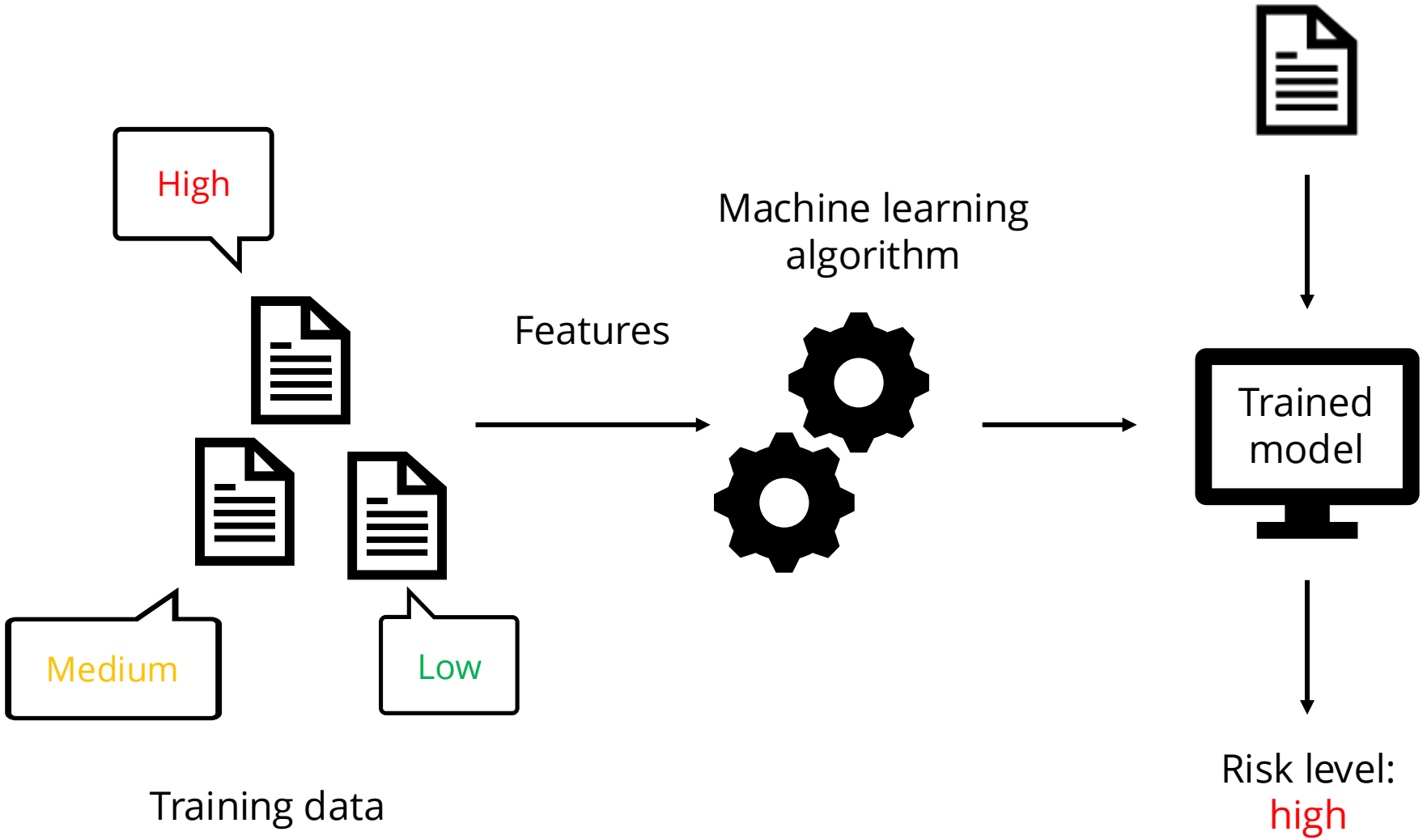
- Filter messages and/or cases, threat assessor only examines those with high priority



Mogelijke toepassingen: dreigingsinschattingen

Full replacement

- AI model (trained on earlier assessments) does the entire assessment instead of humans

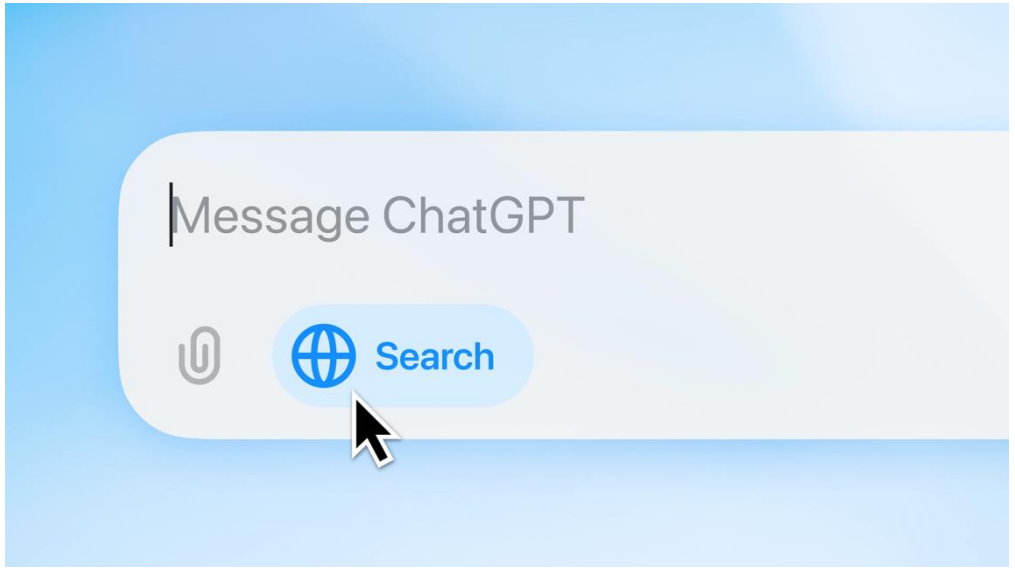

High
Medium
Low
Training data
Features
Machine learning algorithm
Trained model
Risk level:
high

Applications of AI: threat assessment

“Your role as a threat assessor might change completely. Instead of doing the entire analysis yourself, you will only assess or verify AI output. And you might need specific training for that, but your role will be different.”

Applications of AI: threat assessment

- GenAI chat-based ‘assistant’



Challenges

Challenge: Specialized (?) knowledge

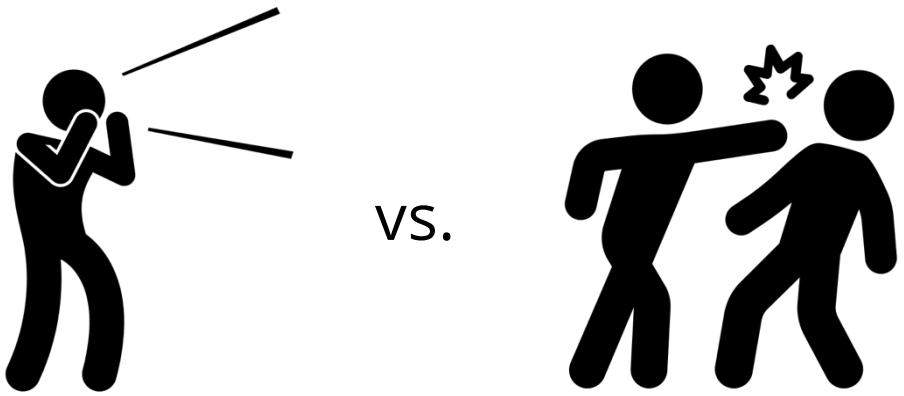
- Nuances of threat assessment and analysis
- ‘Out of the box’ models not trained on real cases

Challenge: Lack of training data

- The behaviour we are trying to detect/predict almost never happens

'There are a lot of awful things said about people online, but is it just another keyboard warrior? Or is it from a dangerous lone actor that we must take seriously? If we were able to develop something to address that, it would be very helpful. But then you will need a lot of cases in which something went wrong, to be able to assess this possibility. And that is what missing.'

Challenge: Lack of training data



Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million documents
- Base rate: 0.01% is violent (10,000)

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence			
	No violence			
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence			10,000
	No violence			
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence	9,900		10,000
	No violence			
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence			
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence			99,990,000
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence		98,990,100	99,990,000
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	**Total**
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence	999,900 (false positives)	98,990,100	99,990,000
	Total			

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)

		Prediction		
		Violence	No violence	Total
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence	999,900 (false positives)	98,990,100	99,990,000
	Total	1,009,800	98,990,200	100,000,000

Challenge: The base rate fallacy

- Imagine: we have 99% accurate system
- 100 million individuals
- Base rate: 0.01% is violent (10,000)
- Just 0.98% of predicted violence are classified correctly (!)

		Prediction		
		Violence	No violence	Total
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence	999,900 (false positives)	98,990,100	99,990,000
	Total	1,009,800	98,990,200	100,000,000

Challenge: The base rate fallacy

- What is the impact of this?

		Prediction		
		Violence	No violence	Total
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence	999,900 (false positives)	98,990,100	99,990,000
	Total	1,009,800	98,990,200	100,000,000

Challenge: Explainability of AI models

- Commercial companies offering AI tools to ‘predict violent behaviour’
- Black box
- Transparency: how and why did an AI system make a decision?

Figure 2: Screenshots of the profile risk assessment tool (PRAT)

Challenge: Explainability of AI models

Impact?

- Commercial companies offering AI tools to ‘predict violent behaviour’
- Black box
- Transparency: how and why did an AI system make a decision?

Figure 2: Screenshots of the profile risk assessment tool (PRAT)

Challenge: Human-AI interaction

- Many unknowns about how AI use will influence human assessments
- Do their assessments actually become better or more efficient? To what extent does it exacerbate bias?

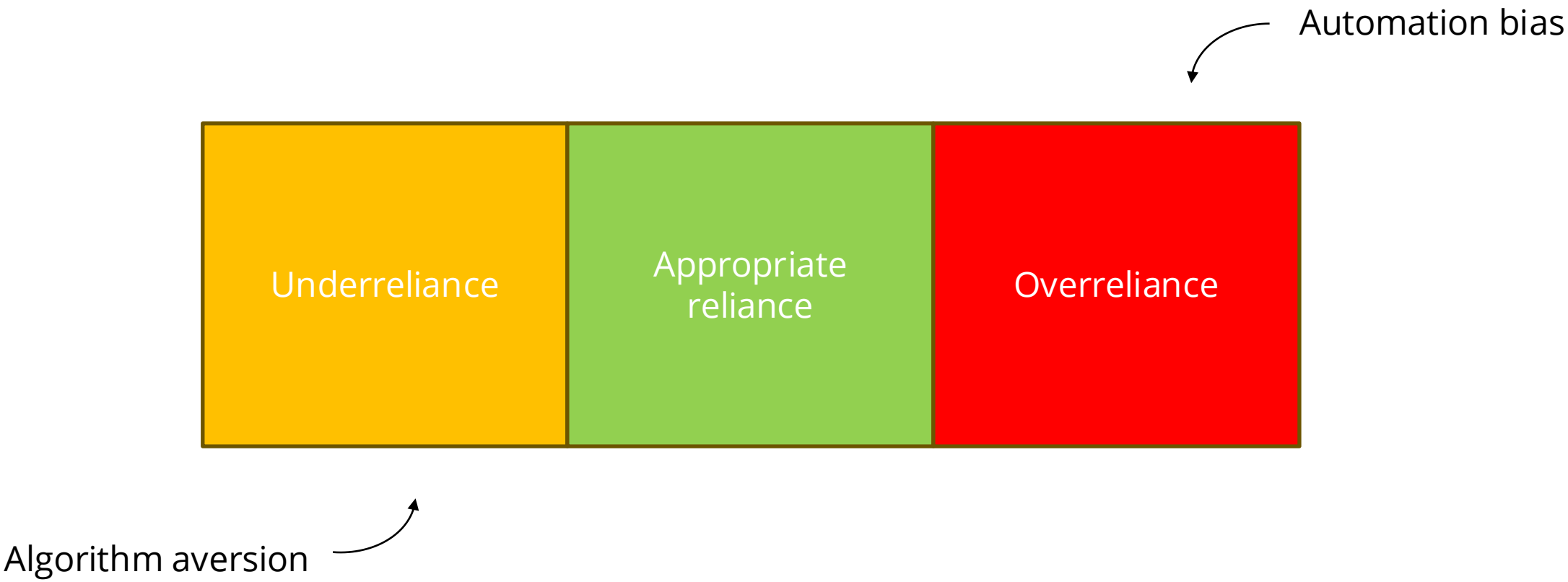
Challenge: Human-AI interaction

- Bias: AI bias more impactful than human bias
- Models trained on earlier biases

Challenge: Human-AI interaction

- Bias: AI bias more impactful than human bias
- Models trained on earlier biases
- Which biases? What impact?

Challenge: Human-AI interaction



Utrecht University

Bruijnes et al., 2025

Challenge: GenAI hallucinations

- LLMs are not trained to produce true or factual information, but to generate plausible text
- LLMs don’t know doubt
- What could be the impact?

Utrecht University

IOS-Themes:
15 platforms

Security in Open Societies

Bottom-up Initiatives for Societal Change

Behaviour and Institutions

Democracy and good governance

Contesting Governance

Futures of Democracy

The Transactional State as an Institution for Good

Future of Work

Open Cities

Transitions and well-being

Fair Transitions

Longtermism and Institutional Change

Markets and Corporations

Openness challenged: the university at risk?

Equality and diversity

In/Equality

Gender, Diversity and Global Justice

Concluding thoughts

- NLP + AI holds a lot of potential for violence prevention, but we must tread carefully

Support Human-in-the-loop Replacement

Concluding thoughts

- No decisions can be made by AI system alone

1. The following AI practices shall be prohibited:

(d) the placing on the market, the putting into service for this specific purpose, or the use of an AI system for making risk assessments of natural persons in order to assess or **predict the risk of a natural person committing a criminal offence,** based solely on the profiling of a natural person or on **assessing their personality traits and characteri**stics; this prohibition shall not apply to AI systems used to **support the human assessment of the involvement of a person in a criminal activity**, which is already based on objective and verifiable facts directly linked to a criminal activity;

Concluding thoughts

- Difficulty of predicting rare events: move away from prediction?
- Risks of LLM usage
- More research is needed to understand actual impact of AI on human assessments

Utrecht University

Sharing science,
shaping tomorrow

i.w.j.vandervegt@uu.nl
isabellevdv.net